DOCUMENT RESUME

ED 367 946 CS 011 623

AUTHOR	Johnson, Eugene G.; And Others
TITLE	Technical Report of the NAEP 1992 Trial State Assessment Program in Reading.
INSTITUTION	Educational Testing Service, Princeton, N.J.; National Assessment of Educational Progress, Princeton, NJ.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	ISBN-0-88685-153-X ISBN-0-16-043109-3; NAEP-23-ST10; NCES-94-472
PUB DATE	Feb 94
NOTE	330p.
AVAILABLE FROM	U.S. Government Printing Office, Superintendent of Documents, Mail Stop: SSOP, Washington, DC 20402-9328 (ISBN-0-16-043109-3).
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC14 Plus Postage.
DESCRIPTORS	Elementary Education; *Evaluation Methods; Grade 4; Grade 8; Program Descriptions; Program Design; Program Implementation; *Reading Achievement; Reading Research; *Research Methodology
IDENTIFIERS	*National Assessment of Educational Progress; *Trial State Assessments (NAEP)

ABSTRACT

Documenting the design and data analysis procedures behind the 1992 Trial State Assessment in reading, this book also provides insight into the rationale behind the technical decisions made about the program. Chapters in the book are: (1) "Overview: The Design, Implementation, and Analysis of the 1992 Trial State Assessment Program in Reading" (John Mazzeo and others); (2) "Developing the Objectives, Cognitive Items, Background Questions, and Assessment Instruments" (Jay R. Campbell and Mary R. Foertsch); (3) "Sample Design and Selection" (Leyla K. Mohadjer and others); (4) "State and School Cooperation and Field Administration" (Nancy Caldwell); (5) "Processing and Scoring Assessment Materials" (Dianne Smrdel and others); (6) "Creation of the Database and Evaluation of the Quality Control of Data Entry" (John J. Ferris and David S. Freund); (7) "Weighting Procedures and Variance Estimation" (Adam Chu and Keith F. Rust); (8) "Theoretical Background and Philosophy of NAEP Scaling Procedures" (Eugene G. Johnson and others); (9) "Data Analysis and Scaling for the 1992 Trial State Assessment in Reading" (Nancy L. Allen and others); and (10) "Conventions Used in Reporting the Results of the 1992 Trial State Assessment in Reading" (John Mazzeo). One hundred nine references, a list of participants in the objectives and item development process; a summary of participation rates; conditioning variables and contrast codings; item response theory parameters for reading items; a description of reporting subgroups; and descriptions of the achievement level setting process and the anchoring process are attached. (RS)

NATIONAL CENTER FOR EDUCATION STATISTICS

Technical Report of the NAEP 1992 Trial State Assessment Program in Reading

ED 367 946

HOW TO...
POETRY
SHORT STORIES
JOURNAL

CS011623

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

THE NATION'S REPORT CARD
naep

Prepared by Educational Testing Service under contract with the National Center for Education Statistics.

Office of Educational Research and Improvement
U.S. Department of Education

BEST COPY AVAILABLE

What is The Nation's Report Card?

THE NATION'S REPORT CARD, the National Assessment of Educational Progress (NAEP), is the only nationally representative and continuing assessment of what America's students know and can do in various subject areas. Since 1969, assessments have been conducted periodically in reading, mathematics, science, writing, history/geography, and other fields. By making objective information on student performance available to policymakers at the national, state, and local levels, NAEP is an integral part of our nation's evaluation of the condition and progress of education. Only information related to academic achievement is collected under this program. NAEP guarantees the privacy of individual students and their families.

NAEP is a congressionally mandated project of the National Center for Education Statistics, the U.S. Department of Education. The Commissioner of Education Statistics is responsible, by law, for carrying out the NAEP project through competitive awards to qualified organizations. NAEP reports directly to the Commissioner, who is also responsible for providing continuing reviews, including validation studies and solicitation of public comment, on NAEP's conduct and usefulness.

In 1988, Congress created the National Assessment Governing Board (NAGB) to formulate policy guidelines for NAEP. The board is responsible for selecting the subject areas to be assessed, which may include adding to those specified by Congress; identifying appropriate achievement goals for each age and grade; developing assessment objectives; developing test specifications; designing the assessment methodology; developing guidelines and standards for data analysis and for reporting and disseminating results; developing standards and procedures for interstate, regional, and national comparisons; improving the form and use of the National Assessment; and ensuring that all items selected for use in the National Assessment are free from racial, cultural, gender, or regional bias.

The National Assessment Governing Board

Mark D. Musick, Chairman
President
Southern Regional Education Board
Atlanta, Georgia

Hon. William T. Randall, Vice Chair
Commissioner of Education
State Department of Education
Denver, Colorado

Parris C. Battle
Education Specialist
Dade County Public Schools
Miami, Florida

Honorable Evan Bayh
Governor of Indiana
Indianapolis, Indiana

Mary R. Blanton
Attorney
Blanton & Blanton
Salisbury, North Carolina

Linda R. Bryant
Dean of Students
Florence Reizenstein Middle School
Pittsburgh, Pennsylvania

Naomi K. Cohen
Office of Policy and Management
State of Connecticut
Hartford, Connecticut

Charlotte Crabtree
Professor
University of California
Los Angeles, California

Chester E. Finn, Jr.
Founding Partner and Senior Scholar
The Edison Project
Washington, DC

Michael J. Guerra
Executive Director
NCEA, Secondary School Department
Washington, District of Columbia

William (Jerry) Hume
Chairman of the Board
Basic American, Inc.
San Francisco, California

Christine Johnson
Director of Urban Initiatives
Education Commission of the States
Denver, Colorado

John S. Lindley
Principal
Galloway Elementary School
Henderson, Nevada

Jan B. Loveless
Jan B. Loveless and Associates
Educational Consultants
Midland, Michigan

Marilyn McConachie
Member, Board of Education
Glenbrook High Schools
Northbrook, Illinois

Honorable Stephen E. Merrill
Governor of New Hampshire
Concord, New Hampshire

Jason Millman
Professor
Cornell University
Ithaca, New York

Honorable Richard P. Mills
Commissioner of Education
State Department of Education
Montpelier, Vermont

John A. Murphy
Superintendent of Schools
Charlotte-Mecklenburg Schools
Charlotte, North Carolina

Mitsugi Nakashima
Hawaii State Board of Education
Honolulu, Hawaii

Michael T. Nettles
Professor
University of Michigan
Ann Arbor, Michigan

Honorable Edgar D. Ross
Senator
Legislature of the Virgin Islands
Christiansted, St. Croix
U.S. Virgin Islands

Marilyn Whirry
English Teacher
Mira Costa High School
Manhattan Beach, California

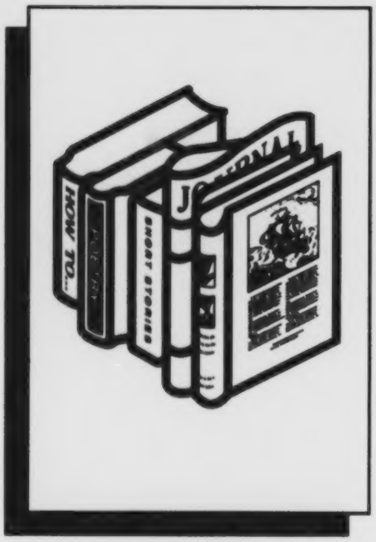
Sharon P. Robinson
Assistant Secretary for Educational
Research and Improvement (Ex-Officio)
U.S. Department of Education
Washington, D.C.

Roy Truby
Executive Director, NAGB
Washington, D.C.

BEST COPY AVAILABLE

NATIONAL CENTER FOR EDUCATION STATISTICS

Technical Report of the NAEP 1992 Trial State Assessment Program in Reading



Eugene G. Johnson
John Mazzeo
Debra L. Kline
in collaboration with

Nancy L. Allen	Robert J. Mislevy
Mary Lyn Bourque	Leyla K. Mohadjer
Drew W. Bowker	Ina V. S. Mullis
Nancy Caldwell	John F. Olson
Jay R. Campbell	Linda Reynolds
Adam Chu	Keith F. Rust
John J. Ferris	Jacqueline Severynse
Mary A. Foertsch	Valerija Smith
Y. Fai Fong	Dianne Smrdel
David S. Freund	Brad Thayer
Steven P. Isham	Neal Thomas

with a Foreword by Gary W. Phillips

Report No. 23-ST10 **February 1994**

CS011623

THE NATION'S REPORT CARD naep

Prepared by Educational Testing Service under contract with the National Center for Education Statistics

Office of Educational Research and Improvement
U.S. Department of Education

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
Sharon P. Robinson
Assistant Secretary

National Center for Education Statistics
Emerson J. Elliott
Commissioner

FOR MORE INFORMATION:

For ordering information on this report, write:

Education Information Branch
Office of Educational Research and Improvement
U.S. Department of Education
555 New Jersey Avenue, NW
Washington, D.C. 20208-5641

or call 1-800-424-1616 (in the Washington, D.C. metropolitan area call 202-219-1651).

Library of Congress, Catalog Card Number: 93-86645

ISBN: 0-88685-153-X

The work upon which this publication is based was performed for the National Center for Education Statistics, Office of Educational Research and Improvement, by Educational Testing Service.

Educational Testing Service is an equal opportunity, affirmative action employer.

Educational Testing Service, ETS, and *the ETS logo* are registered trademarks of Educational Testing Service.

For sale by the U.S. Government Printing Office
Superintendent of Documents, Mail Stop: SSOP, Washington, DC 20402-9328
ISBN 0-16-043109-3

TECHNICAL REPORT
OF THE NAEP 1992 TRIAL STATE ASSESSMENT PROGRAM
IN READING

TABLE OF CONTENTS

List of Tables and Figures ix

Acknowledgments xi

Foreword xv
Gary W. Phillips

Chapter 1 Overview: The Design, Implementation, and Analysis of the 1992 Trial State Assessment Program in Reading 1
John Mazzeo, Eugene G. Johnson, and John F. Olson

1.1 Overview 1
1.2 Design of the Trial State Assessment in Reading 6
1.3 Development of Reading Objectives, Items, and Background Questions 7
1.4 Assessment Instruments 8
1.5 The Sampling Design 9
1.6 Field Administration 10
1.7 Materials Processing and Database Creation 10
1.8 The Trial State Assessment Data 11
1.9 Weighting and Variance Estimation 12
1.10 Preliminary Data Analysis 13
1.11 Scaling the Assessment Items 13
1.12 Linking the Trial State Results to the National Results 14
1.13 Reporting the Trial State Assessment Results 15

Chapter 2 Developing the Objectives, Cognitive Items, Background Questions, and Assessment Instruments 17
Jay R. Campbell and Mary R. Foertsch

2.1 Overview 17
2.2 Framework and Assessment Design Principles 18
2.3 Framework Development Process 19
2.4 Framework for the Assessment 20
2.5 Distribution of Assessment Items 20
2.6 Developing the Cognitive Items 20
2.7 Student Assessment Booklets 26
2.8 Questionnaires 27
  2.8.1 Student Questionnaires 27
  2.8.2 Teacher, School, and Excluded Student Questionnaires 29
2.9 Development of Final Forms 30

Chapter 3 Sample Design and Selection 33
Leyla K. Mohadjer, Keith F. Rust, Valerija Smith, and Jacqueline Severynse

3.1 Introduction and Overview 33
3.2 Sample Selection for the 1991 Field Test 35
  3.2.1 Primary Sampling Units 35
  3.2.2 Selection of Schools and Students 35
  3.2.3 Assignment to Sessions for Different Subjects 36
3.3 Sampling Frame for the 1992 Assessment 36
  3.3.1 Choice of School Sampling Frame 36
  3.3.2 Missing Minority and Urbanization Data 37
  3.3.3 In-scope Schools 39
3.4 Within-state Stratification 39
  3.4.1 Stratification Variables 39
  3.4.2 Urbanization Classification 39
  3.4.3 Minority Classification 51
  3.4.4 Median Household Income 52
  3.4.5 Schools With Fewer Than 20 Students 52
3.5 School Sample Selection for the 1992 Trial State Assessment 53
  3.5.1 Control of Overlap of School Samples for National Educational Studies 53
  3.5.2 Selection of Schools in Small States 55
  3.5.3 States with Geographic Clustering of Small Schools 55
  3.5.4 States with Stratification of Small Schools 55
  3.5.5 Overlap of School Samples 57
  3.5.6 New School Selection 57
  3.5.7 Assigning Subject Session Types at Grade 4 58
  3.5.8 Designating Monitor Status 60
  3.5.9 School Substitution and Participation 60
3.6 Student Sample Selection 61

Chapter 4 State and School Cooperation and Field Administration 67
Nancy Caldwell

4.1 Overview 67
4.2 The Field Test 67
  4.2.1 Conduct of the Field Test 67
  4.2.2 Results of the Field Test 68
4.3 The 1992 Trial State Assessment 69
  4.3.1 Overview of Responsibilities 69
  4.3.2 Schedule of Data Collection Activities 72
  4.3.3 Preparations for the Trial State Assessment 73
  4.3.4 Monitoring of Assessment Activities 75
  4.3.5 School and Student Participation 76
  4.3.6 Results of the Observations 77

Chapter 5 Processing and Scoring Assessment Materials 79
Dianne Smrdel, Linda Reynolds, and Brad Thayer

5.1 Overview 79
5.2 Process Control System 80
5.3 Workflow Management System 80
5.4 Process Flow of NAEP Materials and Database Creation 80
5.5 Materials Distribution 82
5.6 Processing Assessment Materials 84
5.7 Professional Scoring 86
  5.7.1 Description of Scoring 88
  5.7.2 Training 88
  5.7.3 Reliability of Scoring 91
5.8 Data Transcription Systems 92
  5.8.1 Data Entry 92
  5.8.2 Scanning 92
5.9 Data Validation 93
5.10 Editing 94
5.11 Questionnaires 96
5.12 Merging of Student Data 96
5.13 Storage of Documents 97

Chapter 6 Creation of the Database and Evaluation of the Quality Control of Data Entry 99
John J. Ferris and David S. Freund

6.1 Overview 99
6.2 Merging Files into the Trial State Assessment Database 99
6.3 Creating the Master Catalog 100
6.4 Quality Control Evaluation 101
  6.4.1 Student Data 101
  6.4.2 Teacher Questionnaires 103
  6.4.3 School Questionnaires 104
  6.4.4 Excluded Student Questionnaires 104

Chapter 7 Weighting Procedures and Variance Estimation 105
Adam Chu and Keith F. Rust

7.1 Introduction 105
7.2 Calculation of Base Weights 106
  7.2.1 Calculation of School Base Weights 106
  7.2.2 Weighting New Schools 107
  7.2.3 Treatment of Substitute and Double-session Substitute Schools 107
  7.2.4 Calculation of Student Base Weights 108
7.3 Adjustments for Nonresponse 109
  7.3.1 Defining Initial School-level Nonresponse Adjustment Classes 109
  7.3.2 Constructing the Final Nonresponse Adjustment Classes 110
  7.3.3 School Nonresponse Adjustment Factors 111
  7.3.4 Student-level Nonresponse Adjustment Classes 112

7.3.5 Student Nonresponse Adjustments 113
7.4 Characteristics of Nonresponding Schools and Students 114
7.4.1 Weighted Distributions of Schools Before and After Nonresponse 114
7.4.2 Characteristics of Nonresponding Schools 118
7.4.3 Weighted Distributions of Students Before and After Student Absenteeism 122
7.4.4 Characteristics of Absent Students 124
7.5 Variation in Weights 126
7.6 Calculation of Replicate Weights 127
7.6.1 Defining Replicate Groups for Variance Estimation 127
7.6.2 School-level Replicate Weights 129
7.6.3 Student-level Replicate Weights 130
7.7 Calculation of School Weights 131

Chapter 8 Theoretical Background and Philosophy of NAEP Scaling Procedures 133
Eugene G. Johnson, Robert J. Mislevy, and Neal Thomas

8.1 Overview 133
8.2 Background 133
8.3 Scaling Methodology 135
8.3.1 The Scaling Models 135
8.3.2 An Overview of Plausible Values Methodology 139
8.3.3 Computing Plausible Values in IRT-based Scales 141
8.4 Achievement Levels 142
8.5 Analyses 143
8.5.1 Computational Procedures 143
8.5.2 Statistical Tests 144
8.5.3 Biases in Secondary Analyses 145

Chapter 9 Data Analysis and Scaling for the 1992 Trial State Assessment in Reading 147
Nancy L. Allen, John Mazzeo, Steven P. Isham, Y. Fai Fong, and Drew W. Bowker

9.1 Overview 147
9.2 Description of Items, Assessment Booklets, and Administration Procedures 148
9.3 Item Analyses 150
9.3.1 Conventional Item and Test Analyses 150
9.3.2 Differential Item Functioning (DIF) Analyses 153
9.4 Item Response Theory (IRT) Scaling 159
9.4.1 Item Parameter Estimation 161
9.5 Estimation of State and Subgroup Proficiency Distributions 168
9.6 Linking State and National Scales 175
9.7 Producing a Reading Composite Scale 179

Chapter 10 Conventions Used in Reporting the Results of the 1992 Trial State Assessment in Reading 183
John Mazzeo

10.1 Overview 183
10.2 Minimum Sample Sizes for Reporting Subgroup Results 185
10.3 Estimates of Standard Errors with Large Mean Squared Errors 186
10.4 Treatment of Missing Questionnaire Data 187
10.5 Statistical Rules Used for Producing the State Reports 189
  10.5.1 Comparing Means and Proportions for Mutually Exclusive Groups of Students 189
  10.5.2 Multiple Comparison Procedure 190
  10.5.3 Determining the Highest and Lowest Scoring Groups from a Set of Ranked Groups 191
  10.5.4 Statistical Significance and Estimated Effect Sizes 192
  10.5.5 Descriptions of the Magnitude of Percentage 193

* * *

Appendix A Participants in the Objectives and Item Development Process 195

Appendix B Summary of Participation Rates 201

Appendix C Conditioning Variables and Contrast Codings 217

Appendix D IRT Parameters for Reading Items 243

Appendix E Trial State Assessment Reporting Subgroups; Composite and Derived Common Background Variables; Composite and Derived Reporting Variables 249

Appendix F The NAEP Achievement Level Setting Process for the 1992 Reading Assessment 263
Mary Lyn Bourque

Appendix G The NAEP Scale Anchoring Process for the 1992 Reading Assessment 291
Ina V.S. Mullis, Eugene G. Johnson, Jay R. Campbell, and Steven P. Isham

References Cited in Text 309

LIST OF TABLES AND FIGURES

Table 1-1 Jurisdictions participating in the 1992 Trial State Assessment Program 2

Figure 2-1 Description of reading stances 21
2-2 Description of purposes for reading 22
Table 2-1 Percentage distribution of items by grade and reading purpose 23
2-2 Percentage distribution of items by reading stance 23
2-3 Assessment time devoted to reading stances by purpose of reading 26
2-4 Cognitive and noncognitive block information 28
2-5 Booklet contents 28

3-1 Distribution of fourth-grade schools and enrollment as reported in QED 1990 38
3-2 Distribution of the selected schools by sampling strata 40
3-3 Distribution of sample sizes by school size, with corresponding overlap between grades 56
3-4 Distribution of new schools coming from "large" and "small" districts 59
3-5 Substitute school counts 62
3-6 Distribution of the grade 4 reading school sample by state 63
3-7 Distribution of the grade 4 reading student sample and response rates by state 65

Figure 4-1 Participating jurisdictions, 1990 and 1992 Trial State Assessments 70
Table 4-1 School participation, 1992 Trial State Assessment 76
4-2 Student participation, 1992 Trial State Assessment in mathematics 77

Figure 5-1 Data flow overview 81
5-2 Materials processing flow 83
5-3 Packing list 87
Table 5-1 Interreader reliabilities for extended constructed-response items 91
Figure 5-4 Extended constructed-response scoring guide 89

Table 6-1 Number of assessment booklets scanned and selected for quality control evaluation 102
6-2 Inference from the quality control evaluation of student data 103

7-1 Unweighted and weighted counts of assessed students by state 115
7-2 Unweighted and weighted counts of excluded students with returned questionnaires by state 116
7-3 Weighted mean values derived from sampled schools 117
7-4 Grade 4 school nonresponse adjustment classes with adjustment factors greater than 1.25 119
7-5 Weighted student percentages derived from sampled schools 123
7-6 Grade 4 student nonresponse adjustment classes with adjustment factors greater than 1.25 125

9-1 Reading block composition by scale and item type 149
9-2 Descriptive statistics for each block of items by position within test booklet and overall 151
9-3 Block-level descriptive statistics for monitored and unmonitored sessions 154
Figure 9-1 Stem-and-leaf display of state-by-state differences in average item scores by scale (monitored minus unmonitored) 155
Table 9-4 Frequency distributions of DIF statistics for grade 4 items grouped by content area 158
Figure 9-2 Stem-and-leaf display of average item scores by scale 160

Figure 9-3 Differences in item scores (monitored minus unmonitored) plotted against monitored item scores 162
Table 9-5 Extended constructed-response items 164
Figure 9-4 Plots comparing empirical and model-based estimates of item response functions for binary-scored items exhibiting good model fit 166
9-5 Plot comparing empirical and model-based estimates of item category characteristic curves for a polytomously scored item exhibiting good model fit 167
9-6 Plot comparing empirical and model-based estimates of item response functions for binary-scored items exhibiting some model misfit 169
9-7 Plot comparing empirical and model-based estimates of item category characteristic curves for a polytomously score item exhibiting some model misfit 170
9-8 Plot comparing empirical and model-based estimates of the item response function for item R012111 before collapsing unsatisfactory and partial response categories 171
9-9 Plot comparing empirical and model-based estimates of the item response function for item R012111 after collapsing unsatisfactory and partial response categories 172
Table 9-6 Summary statistics for state conditioning models 174
Figure 9-10 Plot of mean proficiency versus mean item score 176
Table 9-7 Transformation constants 178
Figure 9-11 Rootogram comparing proficiency distributions for the trial state assessment aggregate sample and the state aggregate comparison sample from the national assessment for each content area scale 180
Table 9-8 Weights used for each scale to form the reading composite 181
Figure 9-12 Rootogram comparing proficiency distributions for the trial state assessment aggregate sample and the state aggregate comparison sample from the national assessment for the composite scale 182

Table 10-1 Weighted percentage of students matched to reading teacher questionnaire 188
10-2 Rules for selecting descriptions of percentages 193

ACKNOWLEDGMENTS

The design, development, analysis, and reporting of the Trial State Assessment Program was truly a collaborative effort among staff from State Education Agencies, the National Center for Education Statistics (NCES), Educational Testing Service (ETS), Westat, and National Computer Systems (NCS). The program benefitted from the contributions of hundreds of individuals at the state and local levels—Governors, Chief State School Officers, State and District Test Directors, State Coordinators, and district administrators—who tirelessly provided their wisdom, experience, and hard work. Finally, and most importantly, NAEP is grateful to the students and school staff who participated in the Trial State Assessment.

This report documents the design and data analysis procedures behind the 1992 Trial State Assessment in reading. It also provides insight into the rationale behind the technical decisions made about the program. The development of this Technical Report, and especially of the Trial State Assessment Program, is the culmination of effort by many individuals who contributed their considerable knowledge, experience, and creativity to the 1992 Trial State Assessment Program in reading.

The 1992 Trial State Assessment was funded through the National Center of Education Statistics in the Office of Educational Research and Improvement of the U.S. Department of Education. Emerson Elliott, NCES Commissioner, provided consistent support and guidance. The staff—particularly Gary Phillips, Eugene Owen, Stephen Gorman, Peggy Carr, Sharif Shakrani, Susan Ahmed, Andrew Kolstad, and Maureen Treacy—worked closely and collegially with ETS, Westat, and NCS staff and played a crucial role in all aspects of the program.

The members of the National Assessment Governing Board (NAGB) and NAGB staff provided advice and guidance throughout, and their contractor, American College Testing, worked with various panels in setting the achievement levels, and carried out a variety of analyses related to the levels.

NAGB's contractor for the reading consensus project, the Council of Chief State School Officers,worked diligently under tight time constraints to create the forward-thinking framework underlying the assessment.

NAEP owes a great deal to the numerous panelists and consultants who worked so diligently on developing the assessment and providing a frame for interpreting the results, including those who helped create the objectives, develop the assessment instruments, set the achievement levels, and provide the anchoring descriptions.

Under the NAEP contract to ETS, Archie Lapointe served as the executive director and Ina Mullis as the project director. John Barone directed the data analysis activities; Jules Goodison, the operational aspects; Stephen Koffler, test development; Kent Ashworth, information services; Eugene Johnson, measurement and research; and John Olson, technical assistance and state services.

ETS and NAEP management have been very supportive of NAEP's technical work. Special thanks go to Gregory Anrig and Nancy Cole as well as to Henry Braun and Charles Davis of ETS research management, and to Archie Lapointe, Ina Mullis, Jules Goodison, David Hobson, and Paul Williams of NAEP management.

The guidance of the NAEP Design and Analysis Committee on the technical aspects of NAEP has been outstanding. The members are Sylvia Johnson (chair), Albert Beaton, Jeri Benson, John Carroll, Clifford Clogg, William Cooley, Jeremy Finn, Bert Green, Huynh Huynh, David Lohman, Bengt Muthén, Anthony Nitko, Ingram Olkin, Tej Pandey, and Juliet Shaffer.

The design and data analysis of the 1992 Trial State Assessment Program was primarily the responsibility of the NAEP research and data analysis staff, with significant contributions from the NAEP management, Westat, and NCS staffs. Statistical and psychometric activities were led by Nancy Allen and John Donoghue under the direction of Eugene Johnson and John Mazzeo. Major contributions were made by James Carlson, Huahua Chang, Angela Grima, Frank Jenkins, Jo-lin Liang, Eiji Muraki, Spencer Swinton, and Neal Thomas. Robert Mislevy and Ming-mei Wang provided valuable statistical and psychometric advice.

The division of Data Analysis and Technical Research, under the outstanding leadership of John Barone, was responsible for developing the operating systems and carrying out the data analyses. Alfred Rogers and David Freund deserve special recognition for their leadership in developing and maintaining the large and complex NAEP data management systems. Alfred Rogers also deserves special mention for his role in the development of production versions of key analysis and scaling systems. Steven Isham performed the data analyses, assisted by Yim Fai Fong. Special thanks also go to Steven Isham, David Freund, Bruce Kaplan, Edward Kulick, and John J. Ferris for their continuing roles as leaders and developers of innovative software solutions to NAEP data analysis challenges. The individual state-level reports and the Reading Report Card were designed and developed through the superb efforts of Laura Jerry and Robert Patrick, in collaboration with Philip Leung, Bruce Kaplan, John J. Ferris, and Jennifer Nelson. Other members of this division who made substantial contributions of their talent, and important contributions to NAEP data analyses, were Drew Bowker, Laura Jenkins, Michael Narcowich, Craig Pizzuti, Ira Sample, and Minhwei Wang.

The staff of Westat, Inc. contributed their exceptional talents in all areas of sample design and data collection. Particular recognition is due to Renee Slobasky and Nancy Caldwell for supervising the field operations and to Keith Rust for developing and supervising the sampling design. Debra Vivari, Dianne Walsh, Leyla

Mohadjer, Adam Chu, Valerija Smith, and Jacqueline Severynse undertook major roles in these activities also.

Critical to the program was the contribution of National Computer Systems, Inc., which has been responsible for the printing, distribution, and processing of the assessment materials. The leadership roles of John O'Neill and Judith Moyer are especially acknowledged. Thanks go also to Linda Reynolds, Bradley Thayer, Dianne Smrdel, Lavonne Mohn, and Mathilde Kennel.

Judith Alfort, Donna Lembeck, Marciline Yates, and Mary Varone are acknowledged for their patience and diligence in typing and proofing the many revisions of this report.

Kent Ashworth was responsible for coordinating the cover design and final printing of this report.

Special thanks go to Debra Kline for organizing, scheduling, editing, motivating, and ensuring the cohesiveness and correctness of the final report.

Special thanks are also due to many individuals for their invaluable assistance in reviewing the reports, especially the editors who improved the text and the data analysts who checked the accuracy of the data.

FOREWORD

This technical report summarizes some of the most complex statistical methodology used in any survey or testing program in the United States. In its 23-year history, the National Assessment of Educational Progress (NAEP) has pioneered such state-of-the-art techniques as matrix sampling and item response theory models. Today it is the leading survey using the advanced plausible values methodology, which uses a multiple imputation procedure in a psychometric context.

The 1992 Trial State Assessment in reading followed the same basic design as that used for the 1990 and 1992 Trial State Assessments in mathematics. Properties of the 1992 reading assessment common to the 1990 and 1992 mathematics assessment include: 1) continuing the use of focused-BIB spiraling, item response theory models, and plausible values; 2) keeping the national and Trial State Assessment samples separate; 3) doing separate stratifications and conditioning in each of the state samples; 4) making each state sample have power similar to the regional samples from the national assessment (this is how the sample sizes for the states were determined); 5) equating the aggregate of the state samples to the national scale (and doing this via a national subsample that also was representative of the aggregate of the states); 6) limiting the state samples to public schools; and 7) using power rules to determine which subgroup comparisons were supported by sufficient sample sizes (this became the "rule of 62," which was derived from the criterion of needing a sample size large enough to detect an effect size of .50 with a power of .80, given an alpha level of .05 and a design effect of 2).

The 1992 Trial State Assessment provided many opportunities to test the limits of statistical theory and thereby advance the state of the art. Some examples include 1) conditioning on a smaller set of principal components rather than a larger set of background variables and 2) the use of the two-parameter polytomous item response theory model for scaling constructed-response and extended constructed-response items.

The Trial State Assessment has many statistical challenges ahead that must be dealt with. As the project plans for the 1994 assessment, it must find ways to 1) accurately report results for nonpublic schools (which have less well developed sampling frames) and 2) improve the methodology for setting achievement levels.

The NAEP project is not only characterized by elegant statistical procedures, but it is also noted for the dedicated professionalism of its staff. It is the stubborn insistence that surveys are scientific activities and relentless quest for improved methodology that have made NAEP credible for over two decades.

Gary W. Phillips
Associate Commissioner
National Center for Education Statistics

Chapter 1

OVERVIEW:

THE DESIGN, IMPLEMENTATION, AND ANALYSIS OF THE 1992 TRIAL STATE ASSESSMENT PROGRAM IN READING

John Mazzeo, Eugene G. Johnson, and John F. Olson

Educational Testing Service

The National Assessment shall conduct a trial mathematics assessment for the fourth and eighth grades in 1992 and, pursuant to subparagraph (6)(D), shall develop a trial reading assessment to be administered in 1992 for the fourth grade in States which wish to participate, with the purpose of determining whether such an assessment yields valid, reliable State representative data. (Section 406 (i)(2)(C)(i) of the General Education Provisions Act, as amended by Pub. L. 100-297 (20 US.C. 1221e-1(i)(2)(C)(ii)))

1.1 OVERVIEW

In April 1988, Congress reauthorized the National Assessment of Educational Progress (NAEP) and added a new dimension to the program—voluntary state-by-state assessments on a trial basis in 1990 and 1992, in addition to continuing the national assessments that NAEP had conducted since its inception. In this report, we will refer to the voluntary state-by-state assessment program as the Trial State Assessment Program. These assessments, which are designed to provide state representative data, are distinct from the assessment designed to provide nationally representative data, referred to in this report as the national assessment. (This terminology is also used in all other reports of the 1990 and 1992 assessments.) It should be noted that the word trial in Trial State Assessment refers to the Congressionally mandated trial to determine whether such assessments can yield valid, reliable state representative data. All instruments and procedures used in the 1990 and 1992 Trial State and national assessments were previously piloted in field tests conducted in the year prior to the assessment.

The 1990 Trial State Assessment Program collected information on the mathematics knowledge, skills, and understanding of a representative sample of eighth-grade students in public schools in 37 states, the District of Columbia, and two territories. The second phase of the Trial State Assessment Program, conducted in 1992, collected information on the mathematics knowledge, skills, and understanding of a representative sample of fourth- and eighth-grade students and the reading knowledge, skills, and understanding of a representative sample of fourth-grade students in public schools in 44 states, the District of Columbia, and two territories.[1]

Table 1-1 lists the jurisdictions that participated in the 1992 Trial State Assessment Program. More than 100,000 students at grade 4 participated in the reading assessment in those jurisdictions. The students who were assessed in reading were administered the same reading assessment booklets that were used in NAEP's 1992 national grade 4 reading assessment.

Table 1-1
Jurisdictions Participating in the
1992 Trial State Assessment Program

Jurisdictions			
Alabama	Hawaii	*Mississippi**	Pennsylvania
Arizona	Idaho	*Missouri**	Rhode Island
Arkansas	Indiana	Nebraska	*South Carolina**
California	Iowa	New Hampshire	*Tennessee**
Colorado	Kentucky	New Jersey	Texas
Connecticut	Louisiana	New Mexico	*Utah**
Delaware	*Maine**	New York	Virginia
District of Columbia	Maryland	North Carolina	Virgin Islands**
Florida	*Massachusetts**	North Dakota	West Virginia
Georgia	Michigan	Ohio	Wisconsin
Guam	Minnesota	Oklahoma	Wyoming

* These states did not participate in the 1990 Trial State Assessment Program. Illinois, Montana, and Oregon participated in the 1990 program but did not participate in the 1992 program.

** The Virgin Islands participated in the testing portion of the 1992 Trial State Assessment Program. However, in accordance with the legislation providing for participants to review and give permission for release of their results, the Virgin Islands chose not to publish their grade 4 results in the reports.

The reading framework established to guide both the 1992 Trial State Assessment and the 1992 national assessment was developed for NAEP through a consensus project of the Council of Chief State School Officers, funded by the National Assessment Governing Board. In addition, questionnaires completed by the students, their reading teachers, and principals or other school administrators provided an abundance of contextual data within which to interpret the reading results.

---

[1] This report provides the technical details of the 1992 Trial State Assessment in reading. For similar information about the 1992 Trial State Assessment in mathematics, see the *Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics* (Johnson, Mazzeo, & Kline, 1993).

The purpose of this report is to provide the technical information about the 1992 Trial State Assessment in reading. It provides a description of the design for the Trial State Assessment and gives an overview of the steps involved in the implementation of the program from the planning stages through to the analysis and reporting of the data. The report describes in detail the development of the cognitive and background questions, the field procedures, the creation of the database for analysis (from receipt of the assessment materials through scanning, scoring, and creation of the database), and the methods and procedures for sampling, analysis, and reporting. It does not provide the results of the assessment—rather, it provides information on how those results were derived.

Educational Testing Service (ETS) was the contractor for the 1990 and 1992 NAEP programs, including the Trial State Assessment. ETS was responsible for overall management of the programs as well as for development of the overall design, the items and questionnaires, data analysis, and reporting. Westat, Inc., and National Computer Systems (NCS) were subcontractors to ETS. Westat was responsible for all aspects of sampling and of field operations, while NCS was responsible for printing, distribution, and receipt of all assessment materials, and for scanning and professional scoring.

This technical report provides information about the technical bases for a series of reports that have been prepared for the 1992 Trial State Assessment Program in reading, including:

- A *State Report* for each participating jurisdiction that describes the reading proficiency of the fourth-grade public-school students in that jurisdiction and relates their proficiency to contextual information about reading policies and instruction.

- The *NAEP 1992 Reading Report Card for the Nation and the States*, which provides data for all of the jurisdictions that participated in the Trial State Assessment Program as well as the results from the 1992 national reading assessment.

- The *Executive Summary of the NAEP 1992 Reading Report Card for the Nation and the States*, providing the highlights of the *Reading Report Card.*

- The *Data Compendium from the NAEP 1992 Reading Assessment for the Nation and the States*, which includes tables of data relating performance on the reading assessment to a wide variety of demographic, perceptual, and experiential variables.

- *Interpreting NAEP Scales*, which describes past, present, and possible future methods of reporting and interpreting NAEP data. These include percent correct statistics, average percent correct, scale scores, scale anchoring, item mapping, and achievement levels.

- *Data Almanacs* for each jurisdiction that contain a detailed breakdown of the reading proficiency data according to the responses to the student, teacher, and school questionnaires for the population as a whole and for important subgroups of the population. There are six sections to each almanac:

- *The Distribution Data Section* provides percentages of students at or above the three achievement levels and below basic, and percentiles for the scales for each of the standard demographic reporting subgroups.

- *The Student Questionnaire Section* provides a breakdown of the proficiency data according to the students' responses to questions in the three student questionnaires included in the assessment booklets.

- *The Teacher Questionnaire Section* provides a breakdown of the proficiency data according to the teachers' responses to questions in the reading teacher questionnaire.[2]

- *The School Questionnaire Section* provides a breakdown of the proficiency data according to the principals' (or other administrators') responses to questions in the school characteristics and policies questionnaire.

- *The Scale Section* provides a breakdown of selected questions from the questionnaires according to each of the scales measuring areas of reading in the assessment.[3]

- *The Reading Item Section* provides the response data for each reading item in the assessment.

ORGANIZATION OF THE TECHNICAL REPORT

This chapter provides a description of the design for the Trial State Assessment in reading and gives an overview of the steps involved in implementing the program from the planning stages to the analysis and reporting of the data. The chapter summarizes the major components of the program, with references to the appropriate chapters for more details. The organization of this chapter, and of the report, is as follows:

- Section 1.2 provides an overview of the design of the Trial State Assessment Program in reading.

---

[2] Because both mathematics and reading were assessed at the fourth-grade level, the fourth-grade teacher questionnaire asked questions about mathematics and reading programs. The mathematics teachers of the students who participated in the mathematics assessment completed the mathematics questions and the reading teachers of the students in the reading assessment completed the reading questions. All teachers were asked to complete the questions about their educational background and training. For the reading assessment, only the data from the students' reading teachers are included.

[3] Scales for fourth-grade students were created for two purposes of reading: Reading for Literary Experience and Reading to Gain Information.

- Section 1.3 summarizes the development of the reading objectives and the development and review of the items written to measure those objectives. Details are provided in Chapter 2.

- Section 1.4 discusses the assignment of the cognitive and background questions to assessment booklets. An initial discussion is provided of the partially-balanced-incomplete-block (PBIB) spiral design that was used to assign cognitive questions to assessment booklets and assessment booklets to individuals. A more complete description is provided in Chapter 2.

- Section 1.5 outlines the sampling design used for the 1992 Trial State Assessment Program. A fuller description is provided in Chapter 3.

- Section 1.6 summarizes Westat's field administration procedures, including securing school cooperation, training administrators, administering the assessment, and conducting quality control. Further details appear in Chapter 4.

- Section 1.7 describes the flow of the data from their receipt at National Computer Systems through data entry, professional scoring, and entry into the ETS/NAEP database for analysis. Chapters 5 and 6 provide a detailed description of the process.

- Section 1.8 provides an overview of the data obtained from the 1992 Trial State Assessment in reading.

- Section 1.9 summarizes the procedures used to weight the assessment data and to obtain estimates of the sampling variability of subpopulation estimates. Chapter 7 provides a full description of the weighting and variance estimation procedures.

- Section 1.10 describes the initial analyses performed to verify the quality of the data in preparation for more refined analyses, with details given in Chapter 9.

- Section 1.11 describes the item response theory subscales and the overall reading composite that were created for the primary analysis of the Trial State Assessment data. Further discussion of the theory and philosophy of the scaling technology appears in Chapter 8 with details of the scaling process in Chapter 9.

- Section 1.12 provides an overview of the linking of the scaled results from the Trial State Assessment to those from the national reading assessment. Details of the linking process appear in Chapter 9.

- Section 1.13 describes the reporting of the assessment results with further details supplied in Chapter 10.

- A series of appendices provide a list of the participants in the objectives and item development process, a summary of the participation rates, a list of the conditioning variables, the IRT parameters for the reading items, the reporting subgroups, composite and derived common background and reporting variables, and description of the processes used to define achievement levels and to anchor the reading scales.

1.2 DESIGN OF THE TRIAL STATE ASSESSMENT IN READING

The major aspects of the design for the Trial State Assessment in reading included the following:

- Participation at the jurisdiction level was voluntary.

- Only fourth-grade students in public schools were assessed. Unlike the national NAEP program, students in parochial or other private schools were not included in the Trial State program. A representative sample of schools was selected in each participating state or territory, and students were randomly sampled within schools.

- The fourth-grade reading assessment used for the NAEP Trial State Assessment and the national NAEP program consisted of eight 25-minute blocks of exercises. Each block contained one reading passage and a combination of constructed-response and multiple-choice items. Passages selected for the assessment were drawn from authentic texts that might be found and used by students in real, everyday reading. Whole stories, articles, or sections of textbooks were used, rather than excepts or abridgements. The type of items—constructed-response or multiple-choice—was determined by the nature of the task. In addition, the constructed-response items were of two types: *short constructed-response* items required students to respond to a question in a few words or a few sentences while *extended constructed-response* items required students to respond to a question in a few paragraphs. Each student was given two of the eight passages.

- Background questionnaires given to the students, the students' reading teachers, and the principals or other administrators provided a variety of contextual information. The background questionnaires for the Trial State Assessment were identical to those used in the age 9/grade 4 national assessment.

- A complex form of matrix sampling called a partially balanced incomplete block (PBIB) spiraling design was used. With PBIB spiraling, students in an assessment session received different booklets, which provides for greater reading content coverage than would have been possible had every student been administered the identical set of items, without imposing an undue testing burden on the student.

- The assessment time for each student was approximately 63 minutes. Each assessed student was assigned a reading booklet that contained a 5-minute background questionnaire, followed by two of the eight 25-minute blocks containing reading items, a 5-minute reading background questionnaire, and a 3-minute background questionnaire. Sixteen different booklets were assembled.

- The assessments took place in the five-week period between February 3 and March 6, 1992. One-fourth of the schools in each state were assessed each week throughout the first four weeks; the fifth week was reserved for the scheduling of makeup sessions.

- Data collection, by law, was the responsibility of each participating jurisdiction. Security and uniform assessment administration were high priorities. Extensive training was conducted to assure that the administration of the assessment would be administered under standard, uniform procedures. Fifty percent of the assessment sessions were monitored by the contractor's staff.

1.3 DEVELOPMENT OF READING OBJECTIVES, ITEMS, AND BACKGROUND QUESTIONS

The 1992 Trial State Assessment and national NAEP program in reading were based on a reading framework[4] developed through a national consensus process, which was set forth by law, that calls for "active participation of teachers, curriculum specialists, subject matter specialists, local school administrators, parents, and members of the general public" (Public Law 100-297, Part C, 1988).

The process of developing the framework was carried out in late 1989 and early 1990 under the direction of the National Assessment Governing Board (NAGB), which is responsible for formulating policy for NAEP, including developing assessment objectives and test specifications. To prepare the 1992 reading framework, NAGB awarded a contract to the Council of Chief State School Officers (CCSSO). As the framework was being developed, the project staff continually sought guidance and reaction from a wide range of people in the fields of reading and assessment, from school teachers and administrators, and from state coordinators of reading and reading assessment. After thorough discussion and some amendment, the recommended framework was adopted by NAGB in March 1990.

The 1992 NAEP reading assessment measured three general types of text and reading situations, the first two of which were measured at the fourth grade:

Reading for Literary Experience usually involves the reading of novels, short stories, poems, plays, and essays. In these reading situations, readers explore the human condition and consider interplays among events, emotions, and possibilities. In reading for literary experience, readers are guided by what and how an author might write in a specific genre and by their expectations of how the text will be organized. The readers' orientation when reading for literary experience usually involves looking for how the author explores or uncovers experiences and engaging in vicarious experiences through the text.

Reading to Gain Information usually involves the reading of articles in magazines and newspapers, chapters in textbooks, entries in encyclopedias and catalogues, and entire books on particular topics. The type of prose found in such texts has its own features. To understand it, readers need to be aware of those features. For example, depending upon what they are reading, readers need to know the rules of literary criticism, or historical sequences of cause and effect, or scientific taxonomies. In addition, readers read to gain information for different purposes—for example, to find specific pieces of information when preparing a research project, or to get some general information when glancing through a magazine article. These purposes call for different orientations to text from those in reading for a literary experience because readers are focused specifically on acquiring information.

---

[4] *Reading Framework for the 1992 National Assessment of Educational Progress* (Washington, D.C.: National Assessment Governing Board, U.S. Department of Education, 1992). In addition, questionnaires completed by the students, their reading teacher, and principal or other school administrator provided an abundance of contextual data within which to interpret the reading results.

Reading to Perform a Task usually involves the reading of documents such as bus or train schedules; directions for games, repairs, and classroom or laboratory procedures; tax or insurance forms; recipes; voter registration materials; maps; referenda; consumer warranties; and office memos. When they read to perform tasks, readers must use their expectations of the purposes of the documents and the structure of documents to guide how they select, understand, and apply such information. The readers' orientation in these tasks involves looking for specific information so as to do something. Readers need to be able to apply the information, not simply understand it, as is usually the case in reading to be informed. Furthermore, readers engaging in this type of reading are not likely to savor the style or thought in these texts, as they might in reading for literary experience. Reading to Perform a Task was not measured at grade 4.

All items underwent extensive reviews by specialists in reading, measurement, and bias/sensitivity, as well as reviews by representatives from State Education Agencies. The items were field tested in 1991 on a representative group of students. Based on the results of the field test, items were revised or modified as necessary and then again reviewed for sensitivity, content, and editorial concerns. With the assistance of ETS/NAEP staff and outside reviewers, the Reading Item Development Committee selected the items to include in the 1992 assessment.

Chapter 2 includes specific details about developing the objectives and items for the Trial State Assessment. The details of the professional scoring process are given in Chapter 6.

1.4 ASSESSMENT INSTRUMENTS

The assembly of cognitive items into booklets and their subsequent assignment to assessed students was determined by a PBIB design with spiraled administration. Details of the PBIB design are provided in Chapter 2. In addition to the student assessment booklets, three other instruments provided data relating to the assessment—a reading teacher questionnaire, a school characteristics and policies questionnaire, and an excluded student questionnaire.

The *student assessment booklets* contained five sections and included both cognitive and noncognitive items. In addition to two 25-minute sections of cognitive questions, each booklet included two 5-minute sets of general and reading background questions designed to gather contextual information about students, their experiences in reading, and their attitudes toward the subject, and one 3-minute section of motivation questions designed to gather information about the students' levels of motivation for taking the assessment.

The *teacher questionnaire* was administered to the reading teachers of the fourth-grade students participating in the assessment. The questionnaire consisted of three sections and took approximately 20 minutes to complete. The first section focused on teachers' general background and experience. The second section focused on teachers' background related to reading. The third section focused on classroom information about reading.[5]

The *school characteristics and policies questionnaire* was given to the principal or other administrator in each participating school and took about 15 minutes to complete. The questions asked about the principal's background and experience, school policies, programs, facilities, and the composition and background of the students and teachers.

The *excluded student questionnaire* was completed by the teachers of those students who were selected to participate in the Trial State Assessment sample but who were determined by the school to be ineligible to be assessed because they either had an Individualized Education Plan (IEP) and were not mainstreamed at least 50 percent of the time, or were categorized as Limited English Proficient (LEP). Each excluded student questionnaire took approximately three minutes to complete and asked about the nature of the student's exclusion and the special programs in which the student participated.

1.5 THE SAMPLING DESIGN

The target population for the Trial State Assessment Program in reading consisted of fourth-grade students enrolled in public schools. The representative sample of fourth-grade students assessed in the Trial State Assessment came from about 125 public schools in each jurisdiction, unless a jurisdiction had fewer than 125 schools with a fourth grade, in which case all or almost all schools were asked to participate. The sample in each state was designed to produce aggregate estimates for the state and for selected subpopulations (depending upon the size and distribution of the various subpopulations within the state), and also to enable comparisons to be made, at the state level, between administration with monitoring and without monitoring. The schools were stratified by urbanicity, percentage of Black and Hispanic students enrolled, and median household income.

In most states, up to 30 students were selected from each school, with the aim of providing an initial target sample size of approximately 3,000 students per state. The student sample size of 30 for each school was chosen to ensure that at least 2,000 students participated from each state allowing, for school nonresponse, exclusion of students, inaccuracies in the measures of enrollment, and student absenteeism from the assessment. In states with fewer schools, larger numbers of students per school were often required to ensure target samples of roughly 3,000 students. In certain jurisdictions, all eligible fourth graders were targeted for assessment.

Students within a school were sampled from lists of fourth-grade students. The decisions to exclude students from the assessment were made by school personnel, as they were in the national assessment, and used the same criteria for exclusion (described in section 1.4) that were used in the national assessment. Each excluded student was carefully accounted for to estimate the percentage of the state population deemed unassessable and the reasons for exclusion.

---

[5] The fourth-grade Teacher Questionnaire also included sections that focused on classroom information related to mathematics. The mathematics teachers of students participating in the fourth-grade mathematics assessment completed those sections.

Chapter 3 describes the various aspects of selecting the sample for the 1992 Trial State Assessment—the construction of the school frames, the stratification process, the updating of the school frame with new schools, the actual sample selection, and the sample selection for the field test.

1.6 FIELD ADMINISTRATION

The administration of the 1992 program and the 1991 field test involved a collaborative effort between staff in the participating states and schools and the NAEP contractors, especially Westat, the field administration contractor. The purpose of the field test conducted in 1991 was to try out the items and procedures for the 1992 program.

Each jurisdiction volunteering to participate in the 1991 field test and in the 1992 Trial State Assessment was asked to appoint a state coordinator who became the liaison between NAEP staff and the participating schools. At the school level, an assessment administrator was responsible for preparing for and conducting the assessment session in one or more schools. These individuals were usually school or district staff and were trained by Westat. In addition, Westat hired and trained a supervisor for each state. The state supervisors were responsible for working with the state coordinators and overseeing assessment activities. Westat also hired and trained four to eight quality control monitors in each state to monitor 50 percent of the assessment sessions in 1992. During the field test, the state supervisors monitored all sessions.

Chapter 4 describes the procedures for obtaining cooperation from states and provides details about the field activities for both the field test and 1992 program. Chapter 4 also describes the planning and preparations for the actual administration of the assessment, the training and monitoring of the assessment sessions, and the responsibilities of the state coordinators, state supervisors, assessment administrators, and quality control monitors.

1.7 MATERIALS PROCESSING AND DATABASE CREATION

Upon completion of each assessment session, school personnel shipped the assessment booklets and forms from the field to NAEP subcontractor National Computer Systems for professional scoring, entry into computer files, and checking. The files were then sent to Educational Testing Service for creation of the database. Careful checking assured that all data from the field were received. Chapter 5 describes the printing, distribution, receipt, processing, and final disposition of the 1992 Trial State Assessment materials.

The volume of collected data and the complexity of the Trial State Assessment processing design, with its spiraled distribution of booklets, as well as the concurrent administration of this assessment and the national assessments, required the development and implementation of flexible, innovatively designed processing programs and a sophisticated Process Control System. This system, described in Chapter 5, allowed an integration of data entry and workflow management systems that included carefully planned and delineated editing, quality control, and auditing procedures.

Chapter 5 also describes the data transcription and editing procedures. These procedures resulted in the generation of disk and tape files containing various assessment information, including the sampling weights required to make valid statistical inferences about the population from which the Trial State Assessment sample was drawn. Before any analysis could begin, the data from these files underwent a quality control check at ETS. The files were then merged into a comprehensive, integrated database. Chapter 6 describes the transcribed data files, the procedure of merging them to create the Trial State Assessment database, and the results of the quality control process.

1.8 THE TRIAL STATE ASSESSMENT DATA

The basic information collected from the Trial State Assessment in reading consisted of the responses of the assessed students to 85 reading exercises organized around eight distinct reading passages. To limit the assessment time for each student to about one hour, a partially balanced incomplete block (PBIB) spiral design was used to assign a subset of the full exercise pool to each student. The PBIB design differed slightly from the fully balanced incomplete block (BIB) spiral design used for the 1990 and 1992 Trial State Assessments in mathematics. Both the PBIB and BIB designs are variants of matrix sampling designs. Somewhere between 2,000 and 3,000 students were assessed within each state and the District of Columbia; apart from nonresponse, all fourth-grade students were assessed in Guam and about half of all fourth-grade students were assessed in the Virgin Islands[6].

The full set of reading items was divided into eight unique blocks, each requiring 25 minutes for completion. Four of the blocks contained literary passages; the items accompanying these blocks were designed to assess student abilities in Reading for Literary Experience. The other four blocks were based on informational prose passages (e.g., magazine articles, newspaper articles, textbook chapters, etc.) and the items accompanying these passages were designed to assess student abilities in Reading to Gain Information. Each assessed student received a booklet containing two of the eight blocks according to a design that ensured that each block was administered to a representative sample of students within each jurisdiction. The design also ensured that each Reading for Literary Experience block was paired in exactly one booklet with every other Reading for Literary Experience block. Similarly, each Reading to Gain Information block was paired in exactly one booklet with every other Reading to Gain Information block. Furthermore, each Reading for Literary Experience block was paired in exactly one booklet with one of the Reading for Information blocks. The data also included responses to the background questionnaires (described in section 1.4). Further details on the assembly of cognitive instruments and the data collection design can be found in Chapter 2.

The national data to which the Trial State Assessment results were compared came from nationally representative samples of public-school students in the fourth grade. These samples were a part of the full 1992 national reading assessment, in which nationally representative samples of students in public and private schools from three age cohorts were assessed: students who were either in the fourth grade or 9 years old; students who were either in the eighth grade or 13 years old; and students who were either in the twelfth grade or 17 years old.

---

[6] The remaining half of the Virgin Islands fourth-grade students were assessed for the Trial State Assessment in mathematics.

The assessment instruments used in the Trial State Assessment were also used in the fourth-grade national assessments and were administered using the identical procedures in both assessments. The time of testing for the state assessments (February 3 to March 6, 1992) occurred within the time of testing of the national assessment (January 6 to April 3, 1992). The state assessments differed from the national assessment, however, in one important regard: Westat staff collected the data for the national assessment while, in accordance with the NAEP legislation, data collection activities for the Trial State Assessment were the responsibility of each participating jurisdiction. The data collection activities included ensuring the participation of selected schools and students, assessing students according to standardized procedures, and observing procedures for test security. To provide quality control of the Trial State Assessment, a random half of the administrations within each state was monitored.

1.9 WEIGHTING AND VARIANCE ESTIMATION

A complex sample design was used to select the students to be assessed in each of the participating jurisdictions. The properties of a sample from a complex design are very different from those of a simple random sample in which every student in the target population has an equal chance of selection and in which the observations from different sampled students can be considered to be statistically independent of one another. The properties of the sample from the complex Trial State Assessment design were taken into account in the analysis of the assessment data.

One way that the properties of the sample design were addressed was by using sampling weights to account for the fact that the probabilities of selection were not identical for all students. These weights also included adjustments for nonresponse of students and of schools. All population and subpopulation characteristics based on the Trial State Assessment data used sampling weights in their estimation. Chapter 7 provides details on the computation of these weights.

In addition to deriving appropriate estimates of population characteristics, it is essential to obtain appropriate measures of the degree of uncertainty of those statistics. One component of uncertainty is a result of sampling variability, which measures the dependence of the results on the particular sample of students actually assessed. Because of the effects of cluster selection (schools are selected first, then students are selected within those schools), observations made on different students cannot be assumed to be independent of each other (and, in fact, are generally positively correlated). As a result, classical variance estimation formulas will produce incorrect results. Instead, a variance estimation procedure that takes the characteristics of the sample into account was used for all analyses. This procedure, called jackknife variance estimation, is discussed in Chapter 7.

Jackknife variance estimation provides a reasonable measure of uncertainty for any statistic based on values observed without error. Statistics such as the average proportion of students correctly answering a given question meet this requirement, but other statistics based on estimates of student reading proficiency, such as the average reading proficiency of a subpopulation, do not. Because each student typically responds to relatively few items within a particular purpose of reading (i.e., reading for literary experience or reading for information), there exists a nontrivial amount of imprecision in the measurement of the proficiency of a given student. This imprecision adds an additional component of variability to statistics based on estimates of individual proficiencies. The estimation of this component of variability is discussed in Chapter 8.

1.10 PRELIMINARY DATA ANALYSIS

Immediately after receipt from NCS of the computer files containing students' responses, all cognitive and noncognitive items were subjected to an extensive item analysis to assure that each item represented what it was purported to measure.

Each block of cognitive items was subjected to item analysis routines, which yielded for each item the number of respondents, the percentage of responses in each category (100 x item score), the percentage who omitted the item, the percentage who did not reach the item, and the correlation between the item score and the item block score (r-polyserial). In addition, the item analysis program provided summary statistics for each block, including reliability (internal consistency). These analyses were used to check on the scoring of the items, to verify the appropriateness of the difficulty level of the items, and to check for speededness. The results also were reviewed by knowledgeable project staff in search of anomalies that might signal unusual results or errors in the database.

Tables of the weighted percentages of students with responses in each category of each cognitive and background item were created and distributed to each state and jurisdiction. Additional analyses comparing the data from the monitored sessions with those from the unmonitored sessions were conducted to determine the comparability of the assessment data from the two types of administrations. Differential item functioning (DIF) analyses were carried out to identify items that were differentially difficult for various subgroups and to reexamine such items with respect to their fairness and their appropriateness for inclusion in the scaling process. Further details of the preliminary analyses appear in Chapter 9.

1.11 SCALING THE ASSESSMENT ITEMS

The primary analysis and reporting of the results from the Trial State Assessment used item response theory (IRT) scale-score models. Scaling models quantify a respondent's tendency to provide correct answers to the items contributing to a scale as a function of a parameter called proficiency that can be viewed as a summary measure of performance across all items entering into the scale. Three distinct IRT models were used for scaling: 1) 3-parameter logistic models for multiple-choice items; 2) 2-parameter logistic models for short constructed-response items that were scored correct or incorrect; and 3) generalized partial credit models for extended constructed-response items that were scored on a multipoint scale. Chapter 8 provides an overview of the scaling models used. Further details on the application of these models are provided in Chapter 9.

Two distinct scales were created for the Trial State Assessment to summarize fourth-grade students' reading. These scales were defined identically to those used for the scaling of the national NAEP fourth-grade reading data. For grade 4, two purposes-of-reading scales were created: Reading for Literary Experience and Reading for Information. Although the items comprising each scale were identical to those used for the national program, the item parameters for the Trial State Assessment scales were estimated from the combined data from all jurisdictions participating in the Trial State Assessment. Item parameter estimation was based on an item calibration sample consisting of an approximately 25 percent sample of all available data. To ensure equal representation in the scaling process, each jurisdiction was equally represented in the item calibration sample, as were monitored and unmonitored administrations from each jurisdiction. Chapter 9 provides further details about item parameter estimation.

The fit of the IRT model to the observed data was examined within each scale by comparing the estimates of the empirical item characteristic functions with the theoretic curves. For binary-scored items, nonmodel-based estimates of the expected proportions of correct responses to each item for students with various levels of scale proficiency were compared with the fitted item response curve; for the extended constructed-response items, the comparisons were based on the expected proportions of students with various levels of scale proficiency who achieved each score level. In general, the item level results were well fit by the scaling models.

Using the item parameter estimates, estimates of various population statistics were obtained for each jurisdiction. The NAEP methods use random draws ("plausible values") from estimated proficiency distributions for each student to compute population statistics. Plausible values are not optimal estimates of individual student proficiencies; instead, they serve as intermediate values to be used in estimating population characteristics. Under the assumptions of the scaling models, these population estimates will be consistent, in the sense that the estimates approach the model based population values as the sample size increases, which would not be the case for subpopulation estimates obtained by aggregating optimal estimates of individual proficiency. Chapter 8 provides further details on the computation and use of plausible values.

In addition to the plausible values for each scale, a composite of the purposes-of-reading scales was created as a measure of overall reading proficiency. This composite was a weighted average of the two purposes-of-reading scale plausible values, in which the weights were proportional to the relative importance assigned to each purpose in the reading objectives. The definition of the composite for the Trial State Assessment program was identical to that used for the national fourth-grade reading assessment. More details about composite scores may be found in Chapter 9, section 9.7.

1.12 LINKING THE TRIAL STATE RESULTS TO THE NATIONAL RESULTS

The results from the Trial State Assessment were linked to those from the national NAEP through linking functions determined by comparing the results for the aggregate of all students assessed in the Trial State Assessment with the results for fourth-grade students in the State Aggregate Comparison (SAC) subsample of the national NAEP. The SAC subsample of the national NAEP is a representative sample of the population of all grade-eligible public-school students within the aggregate of 41 participating states and the District of Columbia (Guam and the Virgin Islands were not included in the aggregate). Specifically, the grade 4 SAC subsample consists of all fourth-grade students in public schools in the states and the District of Columbia who were assessed in the national cross-sectional reading assessment.

A linear transformation within each scale was used to link the results of the Trial State Assessment to the national NAEP. The adequacy of linear linking was evaluated by comparing, for each scale, the distribution of reading proficiency based on the aggregation of all assessed students at each grade from the participating states and the District of Columbia with the equivalent distribution based on the students in the SAC subsample. In the estimation of these distributions, the students were weighted to represent the target population of public-school students in the specified grade in the aggregation of the states and the District of Columbia. If a linear linking is adequate, the distribution for the aggregate of states and the District of Columbia and that for the SAC subsample would have, to a close approximation, the same shape in terms of the skewness, kurtosis, and higher moments of the distributions. The only differences in the distributions allowed by linear linking would be in the means and variances. To a large degree this was found to be the case.

Each reading scale was linked by matching the mean and standard deviation of the scale proficiencies across all students in the Trial State Assessment (excluding Guam and the Virgin Islands) to the corresponding scale mean and standard deviation across all students in the SAC subsample. Further details of the linking are given in Chapter 9.

1.13 REPORTING THE TRIAL STATE ASSESSMENT RESULTS

Each jurisdiction that participated in the Trial State Assessment received a summary report that provided the state's results with accompanying text and tables, and including national and regional comparisons. These reports were generated by a computerized report-generation system in which graphic designers, statisticians, data analysts, and report writers collaborated to develop shells of the reports in advance of the analysis. These prototype reports were provided to State Education Agency personnel for their reviews and comments. The results of the data analysis were then automatically incorporated into the reports that displayed tables and graphs of the results and interpretations of those results, including indications of subpopulation comparisons of statistical and substantive significance.

Each report contained state-level estimates of mean proficiencies, both for the state as a whole and for categories of the key reporting variables: gender, race/ethnicity, level of parental education, and community type. Results were presented for each scale, for the overall reading composite, and by achievement levels. Results were also reported for a variety of other subpopulations based on variables derived from the student, teacher, and school questionnaires. Standard errors were included for all statistics.

A second report, the *NAEP 1992 Reading Report Card for the Nation and the States*, highlights key assessment results for the nation and summarizes results across the states and territories participating in the assessment. This report contains composite scale results (proficiency means, proportions at or above achievement levels, etc.) for the nation, for each of the four regions of the country, and for each jurisdiction participating in the Trial State

Assessment, both overall and by the primary reporting variables. In addition, overall results are reported for each of the reading scales.

The third report is entitled *Data Compendium from the NAEP 1992 Reading Assessment for the Nation and the States*. Like the *Report Card*, the *Compendium* reports results for the nation and for all of the states and territories participating in the Trial State Assessment. The *Compendium* contains most of the tables included in the *Report Card* plus additional tables that provide composite scale results for a large number of secondary reporting variables.

The fourth report is a six-section almanac. The first section, or "distribution" section, provides results for the achievement levels and percentiles. Three of the sections of the almanac (referred to as proficiency sections) present analyses based on responses to each of the questionnaires (student, reading teacher, and school) administered as part of the Trial State Assessment. The fifth section of the almanac, the scale section, reports proficiency means and associated standard errors for the two purposes-of-reading scales. Results in this section are also reported for the total group in each state, as well as for select subgroups of interest. The final section of the almanac, the "p-value" section, provides the total-group proportion of correct responses to each cognitive item included in the assessment.

The production of the state reports, *Reading Report Card, Data Compendium*, and the almanacs required a large number of decisions about a variety of data analysis and statistical issues. For example, because the demographic characteristics of the fourth-grade public-school students vary widely by state, the proportions of students in the various categories of the race/ethnicity, parental education, and type of community variables varied by state. Chapter 10 documents the major conventions and statistical procedures used in generating the state reports, *Reading Report Card, Data Compendium*, and the almanacs. The chapter describes the rules, based on effect size and sample size considerations, that were used to establish whether a particular category contained data sufficient to report reliable results for a particular state. Chapter 10 also describes the multiple comparison and effect size-based inferential rules that were used for evaluating the statistical and substantive significance of subpopulation comparisons.

To provide information about the generalizability of the results, a variety of information about participation rates was reported for each jurisdiction. This information included school participation rates, both in terms of the initially selected samples of schools and in terms of the finally achieved samples, including replacement schools. The student participation rates, the rates of students excluded due to Limited English Proficiency (LEP) and Individualized Education Plan (IEP) status, and the estimated proportions of assessed students who are classified as IEP or LEP were also reported by state.

Chapter 2

DEVELOPING THE OBJECTIVES, COGNITIVE ITEMS, BACKGROUND QUESTIONS, AND ASSESSMENT INSTRUMENTS

Jay R. Campbell and Mary A. Foertsch

Educational Testing Service

2.1 OVERVIEW

Similar to all previous NAEP assessments, the objectives for the 1992 Trial State Assessment in reading were developed through a broad-based consensus process. To prepare the framework and objectives for the 1992 reading assessment, the National Assessment Governing Board (NAGB) contracted with the Council of Chief State School Officers (CCSSO). The development process involved a steering committee, a planning committee, and CCSSO project staff. Educators, scholars, and citizens, representative of many diverse constituencies and points of view, participated in the national consensus process to design objectives for the reading assessment.

After careful reviews of the objectives, assessment passages were selected and items were developed that were appropriate to those objectives. All items underwent extensive reviews by specialists in reading, measurement, and bias/sensitivity, as well as reviews by state representatives.

The objective and item development efforts were governed by four major considerations:

- As is the case for other NAEP assessments, the objectives for the reading assessment had to be developed through a consensus process, involving subject-matter experts, school administrators, teachers, and parents.
- As outlined in the ETS proposal for the administration of the NAEP contract, the development of the items had to be guided by a Reading Instrument Development Panel and receive further review by state representatives and classroom teachers from across the country. In addition, the items had to be carefully reviewed for potential bias.
- As described in the ETS Standards for Quality and Fairness (ETS, 1987), all materials developed at ETS had to be in compliance with specified procedures.
- As per federal regulations, all NAEP cognitive and background items had to be submitted to a federal clearance process.

This chapter includes details about developing the objectives and items for the Trial State Assessment in reading. The chapter also describes the instruments—the student assessment booklets, reading teacher questionnaire, school characteristics and policies questionnaire, and excluded student questionnaire. Various committees worked on the development of the framework, objectives, and items for the reading assessment. A list of the committees and consultants who participated in the 1992 development process is provided in Appendix A.

2.2 FRAMEWORK AND ASSESSMENT DESIGN PRINCIPLES

The reading objectives framework was designed to focus on reading processes and outcomes, rather than reflect a particular instructional or theoretical approach. It was stated that the framework should focus not on the specific reading skills that lead to outcomes, but rather on the quality of the outcomes themselves. The framework was intended to embody a broad view of reading by addressing the increasing level of literacy needed for employability, personal development, and citizenship. The framework also specified a reliance on contemporary reading research and the use of nontraditional assessment formats that more closely resemble desired classroom activities.

The objectives development was guided by the consideration that the assessment should reflect many of the states' curricular emphases and objectives in addition to what various scholars, practitioners, and interested citizens believed should be included in the curriculum. Accordingly, the committee gave attention to several frames of reference.

- The purpose of the NAEP reading assessment is to provide information about the progress and achievement of students in general rather than to test individual students' ability. NAEP is designed to inform policymakers and the public about reading ability in the United States. Furthermore, NAEP state data can be used to inform states of their students' relative strengths and weaknesses.

- The term "reading literacy" should be used in the broad sense of knowing when to read, how to read, and how to reflect on what has been read. It represents a complex, interactive process that goes beyond basic or functional literacy.

- The reading assessment should use valid and authentic tasks that are both broad and complete in their coverage of important reading behaviors so that the test will be useful and valid, and will demonstrate a close link to desired classroom instruction.

- Every effort should be made to make the best use of available methodology and resources in driving assessment capabilities forward. New types of items and new methods of analysis were recommended for the 1992 NAEP in reading.

- Every effort must be made in developing the assessment to represent a variety of opinions, perspectives, and emphases among professionals, as well as state and local school districts.

2.3 FRAMEWORK DEVELOPMENT PROCESS

The National Assessment Governing Board is responsible for guiding NAEP, including the development of the reading assessment objectives and test specifications. Appointed by the Secretary of Education from lists of nominees proposed by the board itself in various statutory categories, the 24-member board is composed of state, local, and federal officials, as well as educators and members of the public.

NAGB began the development process for the 1992 reading objectives by conducting a widespread mail review of the objectives for the 1990 reading assessment and by holding a series of public hearings throughout the country. The contract for managing the remainder of the consensus process was awarded to the Council of Chief State School Officers. The development process included the following activities:

- A Steering Committee consisting of members recommended by each of 15 national organizations (see Appendix A) was established to provide guidance for the consensus process. The committee responded to the progress of the project and offered advice. Drafts of each version of the document were sent to members of the committee for review and reaction.

- A Planning Committee (see Appendix A) was established to identify the objectives to be assessed in reading in 1992 and prepare the framework document. The members of this committee consisted of experts in reading, including college professors, an academic dean, a classroom teacher, a school administrator, state level assessment and reading specialists, and a representative of the business community. This committee met with the Steering Committee and as a separate group. A subgroup also met to develop item specifications. Between meetings, members of the committee provided information and reactions to drafts of the framework.

- The project staff at the Council of Chief State School Officers met regularly with staff from the National Assessment Governing Board and the National Center for Education Statistics to discuss progress made by the Steering and Planning committees.

During this development process, input and reactions were continually sought from a wide range of members of the reading field, experts in assessment, school administrators, and state staff in reading assessment. In particular, the process was informed by innovative state assessment efforts and work being done by the Center for the Learning and Teaching of Literature (Langer, 1989, 1990).

2.4 FRAMEWORK FOR THE ASSESSMENT

The framework adopted for the 1992 reading assessment is organized according to a four-by-three matrix of reading *stances* by reading *purposes*. The stances include

- Initial Understanding,
- Developing an Interpretation,
- Personal Reflection and Response, and
- Demonstrating a Critical Stance.

These stances were assessed across three global purposes defined as

- Reading for Literary Experience,
- Reading to Gain Information, and
- Reading to Perform a Task.

Different types of texts were used to assess the various purposes for reading. Students' reading abilities were evaluated in terms of a single purpose for each type of text. At grade 4 only Reading for Literary Experience and Reading to Gain Information were assessed, while all three global purposes were assessed at grades 8 and 12. Figures 2-1 and 2-2 describe the four reading stances and three reading purposes that guided the development of the 1992 Trial State Assessment in reading.

2.5 DISTRIBUTION OF ASSESSMENT ITEMS

For 1992, the Planning Committee was interested in creating an assessment that would be forward-thinking and reflect quality instruction. In recognition that the demands made of readers change as they mature and move through school, it was recommended that the proportion of items had some relation to reading purpose (to perform a task, for literary experience, to gain information). The distribution of items by reading purpose across grade levels is provided in Table 2-1.

Readers use a range of cognitive abilities and assume various stances that should be assessed within each of the reading purposes. While reading, students form an initial understanding of the text and connect ideas within the text to generate interpretations. In addition, they extend and elaborate their understanding by responding to the text personally and critically and by relating ideas in the text to prior experiences or knowledge. Table 2-2 shows the distribution of items by reading stance, as specified in the reading framework, for all three grade levels.

2.6 DEVELOPING THE COGNITIVE ITEMS

The development of cognitive items began with a careful selection of grade-appropriate passages for the assessment. Passages were selected from a pool of reading selections contributed by teachers from across the country. The framework stated that the assessment passages should represent authentic, naturally occurring reading material that students may

Figure 2-1

Description of Reading Stances

Readers interact with text in various ways as they use background knowledge and understanding of text to construct, extend, and examine meaning. The NAEP reading assessment framework specified four reading stances to be assessed that represent various interactions between readers and texts. These stances are not meant to describe a hierarchy of skills or abilities. Rather, they are intended to describe behaviors that readers at all developmental levels should exhibit.

Initial Understanding

Initial understanding requires a broad, preliminary construction of an understanding of the text. Questions testing this aspect ask the reader to provide an initial impression or unreflected understanding of what was read. In the 1992 NAEP reading assessment, the first question following a passage was usually one testing initial understanding.

Developing an Interpretation

Developing an interpretation requires the reader to go beyond the initial impression to develop a more complete understanding of what was read. Questions testing this aspect require a more specific understanding of the text and involve linking information across parts of the text as well as focusing on specific information.

Personal Reflection and Response

Personal response requires the reader to connect knowledge from the text more extensively with his or her own personal background knowledge and experience. The focus is on how the text relates to personal experience; questions on this aspect ask the readers to reflect and respond from a personal perspective. For the 1992 NAEP reading assessment, personal response questions were typically formatted as constructed-response items to allow for individual possibilities and varied responses.

Demonstrating a Critical Stance

Demonstrating a critical stance requires the reader to stand apart from the text, consider it, and judge it objectively. Questions on this aspect require the reader to perform a variety of tasks such as critical evaluation, comparing and contrasting, application to practical tasks, and understanding the impact of such text features as irony, humor, and organization. These questions focus on the reader as interpreter/critic and require reflection and judgments.

Figure 2-2

Description of Purposes for Reading

Reading involves an interaction between a specific type of text or written material and a reader, who typically has a purpose for reading that is related to the type of text and the context of the reading situation. The 1992 NAEP reading assessment presented three types of text to students representing each of three reading purposes: literary text for literary experience, informational text to gain information, and documents to perform a task. Students' reading skills were evaluated in terms of a single purpose for each type of text.

Reading for Literary Experience

Reading for literary experience involves reading literary text to explore the human condition, to relate narrative events with personal experiences, and to consider the interplay in the selection among emotions, events, and possibilities. Students in the NAEP reading assessment were provided with a wide variety of literary text, such as short stories, poems, fables, historical fiction, science fiction, and mysteries.

Reading to Gain Information

Reading to gain information involves reading informative passages in order to obtain some general or specific information. This often requires a more utilitarian approach to reading that requires the use of certain reading/thinking strategies different from those used for other purposes. In addition, reading to gain information often involves reading and interpreting adjunct aids such as charts, graphs, maps, and tables that provide supplemental or tangential data. Informational passages in the NAEP reading assessment included biographies, science articles, encyclopedia entries, primary and secondary historical accounts, and newspaper editorials,

Reading to Perform a Task

Reading to perform a task involves reading various types of materials for the purpose of applying the information or directions in completing a specific task. The reader's purpose for gaining meaning extends beyond understanding the text to include the accomplishment of a certain activity. Documents requiring students in the NAEP reading assessment to perform a task included directions for creating a time capsule, a bus schedule, a tax form, and instructions on how to write a letter to a senator. In 1992, reading to perform a task was assessed only at grades 8 and 12.

Table 2-1

Percentage Distribution of Items
by Grade and Reading Purpose

Grade	**Purposes for Reading**		
	Literary Experience	**To Gain Information**	**To Perform a Task**
4	55%	45%	(No Scale)
8	40%	40%	20%
12	35%	45%	20%

Table 2-2

Percentage Distribution of Items
by Reading Stance for Grades 4, 8, and 12

Initial Understanding/ Developing an Interpretation	**Personal Response**	**Critical Stance**
33%	33%	33%

encounter in and out of school. Furthermore, these passages were to be reproduced in test booklets as they had appeared in their original publications. Final passage selections were made by the Reading Instrument Development Panel. Finally, in order to guide the development of items, passages were outlined or mapped to identify essential elements of the text.

The Trial State Assessment included constructed-response (short and extended) and multiple-choice items. The decision to use a specific item type was based on a consideration of the most appropriate format for assessing the particular objective. Both types of constructed-response items were designed to provide an in-depth view of students' ability to read thoughtfully and generate their own responses to reading. Short constructed-response questions, which were scored correct/incorrect, were used when students needed to respond in only one or two sentences in order to demonstrate full comprehension. Extended constructed-response questions, which were scored on a partial credit scale, were used when the task required more thoughtful consideration of the text and engagement in more complex reading processes. Multiple-choice items were used when a straightforward, single correct answer was all that was required. Guided by the NAEP reading framework, the Instrument Development Panel monitored the development of all three types of items to assess objectives in the framework. For more information about item scoring, see section 5.7 of Chapter 5.

The Trial State Assessment at grade 4 included eight different 25-minute "blocks," each consisting of one reading passage and a set of multiple-choice and constructed-response items to assess students' comprehension of the written material. Students were asked to respond to two 25-minute blocks within one booklet.

A carefully developed and proven series of steps were used to create the assessment items that reflected the objectives.

1. Item specifications and prototype items were provided in the 1992 Reading Framework.

2. The Reading Instrument Development Panel provided guidance to NAEP staff about how the objectives could be measured given the realistic constraints of resources and the feasibility of measurement technology. The Panel made recommendations about priorities for the assessment and types of items to be developed.

3. Passages were chosen for the assessment through an extensive selection process that involved the input of teachers from across the country as well as the Reading Instrument Development Panel.

4. Item writers, both inside and outside ETS, with subject-matter expertise and skills and experience in creating items according to specifications, wrote assessment items.

5. The items were reviewed and revised by NAEP/ETS staff and external test specialists.

6. Passages and items were reviewed by grade-appropriate teachers across the country for developmental appropriateness.

7. Representatives from the State Education Agencies met and reviewed all items and background questionnaires (see section 2.8 for a discussion of the background questionnaires).

8. Language editing and sensitivity reviews were conducted according to ETS quality control procedures.

9. Field test materials were prepared, including the materials necessary to secure OMB clearance.

10. The field test was conducted in 23 states, the District of Columbia, and three territories.

11. Representatives from State Education Agencies met and reviewed the field test results.

12. Based on the field test analyses, items for the 1992 assessment were revised, modified and re-edited, where necessary. The items once again under went ETS sensitivity review.

13. The Reading Instrument Development Panel selected the blocks to include in the 1992 assessment.

14. After a final review and check to ensure that each assessment booklet and each block met the overall guidelines for the assessment, the booklets were typeset and printed. In total, the items that appeared in the Trial State Assessment underwent 86 separate reviews, including reviews by NAEP/ETS staff, external reviewers, State Education Agency representatives, and federal officials.

The overall pool of items for the trial state assessment consisted of 85 items, including 35 short constructed-response items, 8 extended constructed-response items, and 42 multiple-choice items. Table 2-3 provides the percentage of assessment time devoted to each reading stance within the two purposes for reading.

Table 2-3
Assessment Time Devoted to the Reading Stances
Within Each Purpose for Reading for the 1992 Reading Trial State Assessment

Grade	Purpose for Reading	Initial Understanding/ Developing an Interpretation	Personal Response	Critical Stance
4	Literary	33%	22%	45%
	Informational	45%	33%	22%
	Total	39%	27%	34%

2.7 STUDENT ASSESSMENT BOOKLETS

Each student assessment booklet included two sections of cognitive reading items and three sections of background questions. The assembly of reading blocks into booklets and their subsequent assignment to sampled students was determined by a *partially balanced incomplete block* (PBIB) design with *spiraled* administration.

The first step in implementing PBIB spiraling for the grade 4 reading assessment required constructing blocks of passages and items that required 25 minutes to complete. These blocks were then assembled into booklets containing two 5-minute background sections, one 3-minute background section, and two 25-minute blocks of reading passages and items according to a partially balanced incomplete block design. The overall assessment time for each student was approximately 63 minutes.

At the fourth-grade level, the blocks measured two purposes for reading—literary and informational. The reading blocks were assigned to booklets in such a way that every block within a given purpose for reading was paired with every other block measuring the same purpose but was only paired with one block measuring the other purpose for reading. Every block appears in four booklets—three times within booklets measuring the same purpose and once in a booklet measuring both purposes. This is the *partially balanced* part of the balanced incomplete block design.

The PBIB design for the 1992 national reading assessment (and also for the trial state assessment) was *focused*—each block was paired with every other reading block assessing the same purpose for reading but not with all the blocks assessing the other purpose for reading. The *focused*-PBIB design also balances the order of presentation of the blocks of items—every block appears as the first cognitive block in two booklets and as the second cognitive block in two other booklets. This design allows for some control of context effects (see Chapter 9).

The design used in 1992 required that eight blocks of grade 4 reading items be assembled into sixteen booklets. The assessment booklets were then *spiraled* and bundled. Spiraling involves interweaving the booklets in a systematic sequence so that each booklet appears an appropriate number of times in the sample. The bundles were designed so that each booklet would appear equally often in each position in a bundle.

The final step in the PBIB-spiraling procedure was the assigning of the booklets to the assessed students. The students within an assessment session were assigned booklets in the order in which the booklets were bundled. Thus, students in an assessment session received different booklets, and only a few students in a session received the same booklet. Across all jurisdictions in the Trial State Assessment, representative and randomly equivalent samples of about 27,650 students responded to each item.

Table 2-4 provides the composition of each block of items administered in the Trial State Assessment Program in reading. Table 2-5 shows the order of the blocks in each booklet and how the 8 cognitive blocks were arranged across the 16 booklets to achieve the PBIB-spiral design.

2.8 QUESTIONNAIRES

As part of the Trial State Assessment (as well as the national assessment), a series of questionnaires was administered to students, teachers, and school administrators. Similar to the development of the cognitive items, the development of the policy issues and questionnaire items was a consensual process that involved staff work, field testing, and review by external advisory groups. A Background Questionnaire Panel drafted a set of policy issues and made recommendations regarding the design of the questions. They were particularly interested in capitalizing on the unique properties of NAEP and not duplicating other surveys (e.g., the National Survey of Public and Private School Teachers and Administrators, the School and Staffing Study, and the National Educational Longitudinal Study).

The Panel recommended a focused study that addressed the relationship between student achievement and instructional practices. The policy issues, items, and field test results were reviewed by the group of external consultants who identified specific items to be included in the final questionnaires. In addition, the Reading Instrument Development Panel and state representatives were consulted on the appropriateness of issues addressed in the questionnaires as they relate to reading instruction and achievement. The items underwent internal ETS review procedures to ensure fairness and quality and were then assembled into questionnaires.

2.8.1 Student Questionnaires

In addition to the cognitive questions, the 1992 Trial State Assessment included two five-minute sets of general and reading background questions designed to gather contextual information about students, their instructional and recreational experiences in reading, and their attitudes toward reading. A one-minute questionnaire was given to students at the end of each booklet to determine students' motivation in completing the assessment and their familiarity with assessment tasks. In order to ensure that all fourth-grade students understood the questions and had every opportunity to respond to them, the three sets of student questionnaires were read aloud by administrators as students read along and responded in their booklets.

The **student demographics (common core) questionnaire** (20 questions) included questions about race/ethnicity, language spoken in the home, mother's and father's level of

Table 2-4

Cognitive and Noncognitive Block Information

Block	Type	Total Number of Items	Number of Multiple-choice Items	Number of Constructed-response Items	Booklets Containing Block
B1	Common Background	20	20	0	30 - 45
R2	Reading Background	14	14	0	30 - 45
RB	Reading Motivation	5	5	0	30 - 45
R3	Reading for Literary Experience	11	6	5	30, 31, 35, 43
R4	Reading for Literary Experience	12	5	7	30, 33, 34, 42
R5	Reading for Literary Experience	11	7	4	31, 32, 34, 44
R6	Reading to Gain Information	10	5	5	36, 39, 40, 44
R7	Reading to Gain Information	10	4	6	37, 38, 40, 42
R8	Reading to Gain Information	10	5	5	38, 39, 41, 43
R9	Reading for Literary Experience	9	4	5	32, 33, 35, 45
R10	Reading to Gain Information	12	6	6	36, 37, 41, 45

Table 2-5

Booklet Contents

Booklet Number	Common Background Block	Cognitive Blocks		Reading Background Block	Reading Motivation Block
		1st	2nd		
R30	B1	R4	R3	R2	RB
R31	B1	R3	R5	R2	RB
R32	B1	R5	R9	R2	RB
R33	B1	R9	R4	R2	RB
R34	B1	R4	R5	R2	RB
R35	B1	R3	R9	R2	RB
R36	B1	R6	R10	R2	RB
R37	B1	R10	R7	R2	RB
R38	B1	R7	R8	R2	RB
R39	B1	R8	R6	R2	RB
R40	B1	R6	R7	R2	RB
R41	B1	R10	R8	R2	RB
R42	B1	R7	R4	R2	RB
R43	B1	R8	R3	R2	RB
R44	B1	R5	R6	R2	RB
R45	B1	R9	R10	R2	RB

education, reading materials in the home, homework, attendance, which parents live at home, and which parents work. This questionnaire was the first section in every booklet. In many cases the questions used were continued from prior assessments.

Three categories of information were represented in the second five-minute section of reading background questions called the **student reading questionnaire** (14 questions):

> *Time Spent Studying Reading*: Time spent on task and reading coursework has been shown to be strongly related to reading achievement (Anderson, Hiebert, Scott, & Wilkinson, 1984). Students were asked to describe both the amount of instruction they received in reading and the time spent on reading homework.
>
> *Instructional Practices*: The nature of students' reading instruction is also thought to be related to achievement (Dole, Duffy, Roehler, & Pearson, 1991). Students were asked to report their instructional experiences related to reading in the classroom, including group work, special projects, and writing in response to reading. In addition, they were asked about the instructional practices of their reading teachers and the extent to which the students themselves discussed what they read in class and demonstrated use of skills and strategies.
>
> *Attitudes Towards Reading*: Students' enjoyment of and confidence in their abilities in reading and their perceptions of the usefulness of reading to their present and future lives appear to be related to reading achievement (Guthrie & Greaney, 1991). Students were asked a series of questions about their attitudes and perceptions about reading, such as whether they enjoyed reading and whether they were good in reading.

The student motivation questionnaire (5 questions) asked students to describe how hard they tried on the NAEP reading assessment, how difficult they found the assessment, how many questions they thought they got right, how important it was for them to do well, and how familiar they were with the assessment format.

2.8.2 Teacher, School, and Excluded Student Questionnaires

To supplement the information on instruction reported by students, the reading teachers of the fourth graders participating in the Trial State Assessment were asked to complete a questionnaire about their instructional practices, teaching backgrounds, and characteristics. The teacher questionnaire contained two parts.[1] The first part pertained to the teachers' background and general training. The second part pertained to specific training in teaching reading and the procedures the teacher uses for *each class* containing an assessed student.

---

[1] Because the Trial State Assessment at grade four included both mathematics and reading, the fourth grade teacher questionnaire contained three sections. The first asked about the teachers' background and training, the second asked about classroom information for the mathematics teachers of the students involved in the mathematics assessment, and the third asked about classroom information for the reading teachers of the students involved in the reading assessment. Reading teachers of students participating in the reading assessment were asked to complete the background and reading classroom parts.

The Teacher Questionnaire, Part I: Background and General Training (23 questions) included questions pertaining to gender, race/ethnicity, years of teaching experience, certification, degrees, major and minor fields of study, coursework in education, coursework in specific subject areas, amount of in-service training, extent of control over instructional issues, and availability of resources for their classroom.

The **Teacher Questionnaire, Part II: Training in Reading and Classroom Instructional Information** (56 questions) included questions on the teacher's exposure to various issues related to reading and teaching reading through pre- and in-service training, ability level of students in the class, whether students were assigned to the class by ability level, time on task, homework assignments, frequency of instructional activities used in class, methods of assessing student progress in reading, instructional emphasis given to the reading abilities covered in the assessment, and use of particular resources.

A **School Characteristics and Policies Questionnaire** was given to the principal or other administrator of each school that participated in the trial state assessment program. This information provided an even broader picture of the instructional context for students' reading achievement. This questionnaire (77 questions) included questions about background and characteristics of school principals, length of school day and year, school enrollment, absenteeism, drop-out rates, size and composition of teaching staff, policies about grouping students, curriculum, testing practices and uses, special priorities and school-wide programs, availability of resources, special services, community services, policies for parental involvement, and school-wide problems.

The **Excluded Student Questionnaire** was completed by the teachers of those students who were selected to participate in the trial state assessment sample but who were determined by the school to be ineligible to be assessed. In order to be excluded from the assessment, students must have had an Individualized Education Plan (IEP) and had not mainstreamed at least 50 percent of the time or were categorized as Limited English Proficient (LEP). In addition, the school staff would have needed to determine that it was inappropriate to include these students in the assessment. This questionnaire asked about the nature of the student's exclusion and the special programs in which the student participated.

2.9 DEVELOPMENT OF FINAL FORMS

The field tests were conducted in February and March 1991 and involved 6,800 students in 233 schools in 23 states, the District of Columbia, and three territories. The intent of the field test was to try out the items and procedures and to give the states and the contractors practice and experience with the proposed materials and procedures. About 500 responses were obtained to each item in the field test.

The field test data were collected, scored, and analyzed in preparation for meetings with the Reading Instrument Development Panel. Using item analysis, which provided the mean percentage of correct responses, the r-biserial correlations, and the difficulty level for each item in the field test, committee members, ETS test development staff, and NAEP/ETS staff reviewed the materials. In addition, another meeting of representatives from state education agencies was convened to review the field test results. Four objectives guided these reviews: to determine which items were most related to achievement; to determine the need for revisions of items that lacked clarity, or had ineffective item formats; to prioritize items to be included in the Trial State Assessment; and to determine appropriate timing for assessment items.

Once the committees had selected the items, all items were rechecked for content, measurement, and sensitivity concerns. The federal clearance process was initiated in June 1991 with the submission of draft materials to NCES. The final package containing the final set of cognitive items assembled into blocks and questionnaires was submitted in August 1991. Throughout the clearance process, revisions were made in accordance with changes required by the government. Upon approval, the blocks (assembled into booklets) and questionnaires were ready for printing in preparation for the assessment.

Chapter 3

SAMPLE DESIGN AND SELECTION

Leyla K. Mohadjer, Keith F. Rust, Valerija Smith, and Jacqueline Severynse

Westat, Inc.

3.1 INTRODUCTION AND OVERVIEW

The 1992 Trial State Assessment Program included assessments in eighth-grade mathematics, fourth-grade mathematics, and fourth-grade reading. For the reading assessment, a representative sample of fourth-grade public-school students was drawn in each participating state or territory. The sample was designed to produce aggregate estimates as well as estimates for various subpopulations with approximately equal precision for each participating state. The sample in each state consisted of about 2,500 fourth-graders from about 100 public schools in each case.

The target populations for the fourth-grade component of the 1992 Trial State Assessment Program included only students in regular public schools[1] who were enrolled in the fourth grade at the time of assessment. The sampling frame included the public schools having fourth grade in each state or territory. The samples were selected based on a two-stage sample design—selection of schools within participating states and selection of students within schools. The first-stage samples of schools were selected with probability proportional to the fourth-grade enrollment in the schools. Special procedures were used for states with many small schools, and for states or territories having a small number of schools (see section 3.4.5).

The sampling frame for each state was first stratified by the urbanization status of the area in which the school was located. The urbanization classes were defined in terms of large or mid-size central city, urban fringe of large or mid-size city, large town, small town, and rural areas (see section 3.4.2). Within urbanization strata, schools were further stratified explicitly on the basis of minority enrollment in those states with substantial Black or Hispanic student populations. Minority enrollment was defined as the total percent of Black and Hispanic students enrolled in a school (see section 3.4.3). Within minority strata, schools were sorted by median household income of the ZIP code area where the school was located (see section 3.4.4).

---

[1] A public school is defined as an institution which provides educational services and has one or more grade groups (PK-12) or which is ungraded, has one or more teachers to give instruction, is located in one or more buildings, has an assigned administrator, receives public funds as primary support, and is operated by an education agency. A regular school is a public elementary/secondary school that does not focus primarily on vocational, special, or alternative education.

One of the goals of the 1992 state sample design was to minimize overlap—between the state and national samples, between the state fourth- and eighth-grade samples (in schools that had both grades), and with the first phase followup to *Prospects: The National Longitudinal Study of Chapter I Children* (Abt Associates, 1991).

Systematic samples (see Cochran, 1977, Chapter 8) of fourth-grade schools were selected with probability proportional to the fourth-grade enrollment of the school from the fourth-grade sampling frames in the participating states. The number of schools drawn for the fourth-grade sample varied by state depending on the distribution of the fourth-grade enrollment in each state (see Table 3-1). In those states and territories that had fewer than 100 schools with fourth grade, all schools were included in the sample.

Successive schools were paired, using the same order in which they were selected, and one member of each pair was designated at random to be monitored during the assessment by Westat field staff so that reliable comparisons could be made between sessions administered with and without monitoring.

Both reading and mathematics sessions were conducted in fourth-grade sampled schools in which there were more than 20 students. Schools that had no more than 20 fourth-grade students were randomly assigned to administer either reading or mathematics. Approximately 2,500 students were assessed for each subject and each grade in a given state. Except in the two small territories, about 5,000 fourth-grade students participated in the assessment. On average, 128 fourth-grade schools were sampled in each state (in which sampling of schools was conducted) with about 115 conducting both mathematics and reading assessments, and about 13 conducting only mathematics or reading. The maximum number of schools selected in a state was 200.

Each selected school provided a list of eligible enrolled students, from which a systematic sample of students was drawn. Generally, 60 students were selected for each school from the grade 4 student lists. If there were fewer than 60 students on a list, all students were selected. Selected students within each of the fourth-grade schools were alternately assigned to either the mathematics or the reading assessments.

The 1992 assessment was preceded in 1991 by a field test, the principal goals of which were to test procedures and new items contemplated for the 1992 assessment. Three states and one territory also used the field test to observe and react to proposed strategies. Twenty-four jurisdictions participated in the field test. Schools that participated in the field test were given a chance of selection in the 1992 assessment, and there was no attempt to control the overlap between the school samples for the 1991 field test and those for the 1992 assessment. Section 3.2 documents the procedures used to select the schools for the field test.

Section 3.3 describes the construction of the sampling frames, including the sources of school data, missing data problems, and definition of in-scope schools. Section 3.4 includes a description of the various steps in stratification of schools within participating states. School sample selection procedures (including new and substitute schools) are described in section 3.5. Section 3.6 includes the steps involved in selection of students within participating schools.

3.2 SAMPLE SELECTION FOR THE 1991 FIELD TEST

The Trial State Assessment 1991 field test was conducted together with the field test for the national portion of the assessment. Twenty-four states participated in the field test, which was conducted for grades 4, 8, and 12. Pairs of schools were identified, with one of each pair to be included in the test. This allowed state participation in the selection of the test schools and also facilitated replacement of schools that declined to participate in the assessment. Sampling weights were not computed for the field test samples.

3.2.1 Primary Sampling Units

The frame of field-test PSUs was derived from the frame of NAEP PSUs[2], splitting PSUs where necessary in such a way that each of the new PSUs was completely contained within a single state. Each state was stratified by urbanization/minority. The sample sizes were assigned in such a way that for each NAEP region the sample sizes were proportional to the population of the participating states. Two PSUs were selected from each state. From each of the state strata, once the sample was assigned, the PSUs were selected with probability proportional to the 1980 population counts. The PSUs selected as noncertainties in the NAEP 1990 national sample were excluded from the PSU frame to avoid undue burden on the schools and districts in these PSUs. Controlled selection of PSUs was used to achieve the selection of two PSUs per state, assigned proportionately among strata within each region.

Since two PSUs were selected for each of the participating states, the sample assignment was not proportional to the population counts. Overall, within each region, the assignment of PSUs was proportional to the urbanization/minority stratum population in each region, where the urbanization/minority stratum population distribution was based only upon the participating states, with each state contributing equally. So, for example, the rural population had disproportionately higher representation in the field test than in the general population, since many of the participating states were relatively rural in nature.

3.2.2 Selection of Schools and Students

Public schools with fourth-grade students were in scope for the reading assessment. Schools with fewer than 25 students per grade were eliminated from the frame, to eliminate the relatively high cost per student of conducting assessments in small schools.

The selection of schools avoided overlap with schools that had been selected from the certainty PSUs for the 1990 NAEP national sample and the IEA Reading Literacy Study, conducted for the National Center for Education Statistics (Rust & Bryant, 1991). Also, there was no overlap among the different grade samples.

---

[2] The frame of NAEP PSUs was the frame used to draw the national NAEP samples for 1986 to 1992. Refer to the 1990 national technical report (Johnson & Allen, 1992) for more information.

From each PSU, a sample of five schools was selected with probability proportional to the fourth-grade enrollment. In the states where one PSU had fewer than five schools, the sample from the other PSU in the state was increased so that the overall state sample was still 10 schools per grade. For each school, where the size of the PSU allowed, the second member of each pair was selected in such a way that the "distance" from the primary selection, based on percent of Black students, percent of Hispanic students, grade enrollment, and percent of students living below the poverty line, was the smallest. The overlap of samples was avoided by first selecting the twelfth-grade sample, then eliminating the selected schools from the eighth-grade sample selection, and then eliminating the twelfth- and eighth-grade selections before selecting the fourth-grade sample.

In the fourth-grade sample, two classrooms were selected randomly from each of the three largest schools and one classroom from each of the remaining seven schools. An exception was made in Florida, Kentucky, and Wisconsin, where 50 students were sampled from each of the three largest schools and 25 students from each of the remaining schools (unless the number of students was fewer than 35, in which case all of the them were taken in the sample). These three states wished to try out the student sampling procedures proposed for the 1992 assessment, and so did not use samples of intact classrooms.

3.2.4 Assignment to Sessions for Different Subjects

Three types of sessions were assigned for the field test: print-administered mathematics, audiotape-administered mathematics, and print-administered reading and writing. At grade 4, one classroom (session) per PSU was selected with equal probability to be administered the print-administered mathematics assessment in all states but Florida, Kentucky, and Wisconsin, for a total of 61 such sessions. The remaining 228 sessions were assigned to reading and writing, from which 15 sessions were selected for audiotaped mathematics sessions with equal probability, after implicitly stratifying by geographic and urbanization/minority characteristics. In Florida, Kentucky, and Wisconsin, where samples of 50 students were drawn from the selected schools, the sample was randomly split in two equal sessions. Half of the sessions were randomly assigned to the print-administered mathematics assessment and the rest to the reading assessment. Florida, Kentucky, Wisconsin, and the Virgin Islands did not participate in any audiotaped mathematics sessions or writing sessions, since those two components were not planned to be part of the 1992 Trial State Assessment.

3.3 SAMPLING FRAME FOR THE 1992 ASSESSMENT

3.3.1 Choice of School Sampling Frame

In order to draw the school samples for the 1992 Trial State Assessment, it was necessary to obtain a comprehensive list of public schools in each state. For each school, useful information for stratification purposes, reliable information about grade span and enrollment, and accurate information for identifying the school to the state coordinator (district membership, name, address) were required.

Based on the experience with the 1990 Trial State Assessment, and national assessments in 1984, 1986, 1988, and 1990, the file made available by Quality Education Data, Inc. (QED) was elected as the sampling frame. The National Center for Education Statistics' Common Core of Data (CCD) school file was used to check the completeness of the QED file. This approach differed from that used to develop frames for the 1990 Trial State Assessment, for which the CCD was used primarily. There were several reasons for this change.

For 1992, it was possible to obtain a version of the QED file that contained all of the relevant variables from the most current CCD file. This meant in particular that data on minority enrollment by school, an important school stratification variable, were available on the QED file. These data had been available only on the CCD for the 1990 assessment. In addition, "type of locale," a seven-level urbanization variable newly created by the National Center for Education Statistics, was available on the QED (as well as the CCD) for 1992. The experience in 1990 indicated that, generally speaking, the updatedness of the school lists and the quality of name and address information was both higher overall and more uniform on the QED. This is important for three reasons: 1) an outdated list leads to the selection of relatively many out-of-scope schools and greater reliance on new school sampling procedures; 2) poor quality name and address information leads to errors in the identification of sample schools by state coordinators (some schools on the CCD in 1990 had no city name as part of the address, for example); and 3) good quality ZIP codes are needed to give good stratification by household income (see section 3.4.4).

Based on combination of the above factors, the QED file was chosen as the basis of the frame for each state. The QED list covers all states and territories except Puerto Rico (which did not participate). The version of the QED file used was released in late 1990, in time for selection of the school sample in early 1991. The file was missing minority and urbanization data for a sizable minority of schools (due to the inability of QED to match these schools with the corresponding CCD file). Considerable efforts were undertaken to obtain these variables for all schools in states where these variables were to be used for stratification. These efforts are described in the next section.

Table 3-1 shows the distribution of fourth-grade schools and enrollment within schools as reported in the 1990 QED file. Enrollment was estimated for each grade as the ratio of total school enrollment by the number of grades in the school. Refer to section 3.4.5 for the definition of small school cluster type. Schools with fewer than 20 fourth-grade students were denoted as small schools.

3.3.2 Missing Minority and Urbanization (Type of Locale) Data

As stated earlier, the sampling frame for the 1992 Trial State Assessment was the most recent version of the QED file, as of January 1991. The CCD file was used to extract information on minority and urbanization in the cases where these variables were missing on the QED file. The minority data were extracted only for those schools in states in which minority stratification was performed. In cases where urbanization could not be determined from the CCD file, the three-level classification of urban/suburban/rural (available for all schools on the QED file) was used to impute for urbanization.

Table 3-1
Distribution of Fourth-grade Schools and Enrollment as Reported in QED 1990

State	Small School Cluster Type	Total Schools	Small Schools	Large Schools	Total Enrollment	Small School Enrollment
Alabama	Geographic	786	29	757	59,127	438
Arizona	Geographic	637	57	580	51,261	505
Arkansas	Geographic	550	40	510	35,107	606
California	Geographic	4,610	299	4,311	383,265	2,830
Colorado	Geographic	752	84	668	45,845	860
Connecticut	Geographic	563	11	552	37,069	163
Delaware	None	54	2	52	6,842	32
District of Columbia	None	118	3	115	6,206	34
Florida	Geographic	1,321	13	1,308	144,789	191
Georgia	Geographic	1,021	11	1,010	94,572	178
Guam	None	21	0	21	2,115	0
Hawaii	Geographic	170	3	167	14,070	21
Idaho	Stratified	304	44	258	18,069	385
Indiana	Geographic	1,167	21	1,146	75,807	339
Iowa	Stratified	794	84	710	37,786	1,236
Kentucky	Geographic	832	51	781	50,856	753
Louisiana	Geographic	788	44	744	62,780	627
Maine	Stratified	405	122	283	16,616	1,358
Maryland	Geographic	755	12	743	54,316	155
Massachusetts	Geographic	1,038	28	1,010	64,274	390
Michigan	Geographic	1,876	62	1,814	123,028	571
Minnesota	Geographic	838	66	772	58,711	956
Mississippi	Geographic	465	3	462	41,063	46
Missouri	Stratified	1,093	147	946	63,555	1,728
Nebraska	Stratified	1,011	615	396	21,834	3,226
New Hampshire	Stratified	268	55	213	13,721	654
New Jersey	Geographic	1,338	42	1,296	84,148	639
New Mexico	Stratified	378	57	321	24,316	673
New York	Geographic	2,259	44	2,215	191,873	565
North Carolina	Geographic	1,109	25	1,084	85,158	361
North Dakota	Stratified	359	180	179	9,973	1,628
Ohio	Geographic	2,039	44	1,995	136,626	651
Oklahoma	Stratified	973	216	757	48,217	2,696
Pennsylvania	Geographic	1,879	47	1,832	126,166	727
Rhode Island	Geographic	177	2	175	11,114	28
South Carolina	Geographic	552	4	548	49,117	50
Tennessee	Geographic	933	66	867	66,932	900
Texas	Geographic	3,053	238	2,815	268,796	2,896
Utah	Geographic	432	31	401	36,629	260
Virginia	Geographic	1,041	39	1,002	80,886	523
Virgin Islands	None	24	1	23	1,874	15
West Virginia	Stratified	637	104	533	25,532	1,474
Wisconsin	Stratified	1,147	128	1,019	59,965	1,910
Wyoming	Stratified	238	91	147	8,050	528

3.3.3 In-scope Schools

The target population for the 1992 fourth-grade Trial State Assessment in reading included students in regular public schools who were enrolled in the fourth grade. Catholic diocesan, other private, Bureau of Indian Affairs, Department of Defense, and special education schools were not included.

3.4 WITHIN-STATE STRATIFICATION

3.4.1 Stratification Variables

Selection of schools within participating states involved three stages of explicit stratification and one stage of implicit stratification. The first three stages were school size (where size was the grade level enrollment of the schools), urbanization, and minority enrollment. The final stage was median income.

The first stage of stratification applied only to states with relatively many students in small schools. These states were known as Cluster Type 3 states. The schools were stratified into two strata, one stratum consisting of schools with 20 or more fourth-grade students, and another stratum consisting of all schools with fewer than 20 students in the fourth grade. The primary purpose of this stratification was to ensure that the sample of schools would provide an appropriate student sample size. It also ensured appropriate representation of small schools in states with any substantial number of such schools. Table 3-2 provides the type of stratification used in each of the participating states or territories for the fourth-grade samples, together with counts of the total number of sampled fourth-grade schools that had eligible students enrolled (in scope).

3.4.2 Urbanization Classification

The NCES "type of locale" variable was used to stratify fourth-grade schools into seven different urbanization levels:

1) *Large Central City:* a central city of a Metropolitan Statistical Area (MSA) with a population greater than or equal to 400,000, or a population density greater than or equal to 6,000 persons per square mile.

2) *Mid-size Central City:* a central city of an MSA but not designated as a large central city.

3) *Urban Fringe of Large City:* a place within an MSA of a large central city and defined as urban by the U.S. Bureau of Census.

4) *Urban Fringe of Mid-size City:* a place within an MSA of a mid-size central city and defined as urban by the U.S. Bureau of Census.

Table 3-2
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	In-scope Schools in Strata
ALABAMA (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	9
Mid-size Central City	Medium Percent Minority	9
Mid-size Central City	High Percent Minority	8
Urban Fringe of Mid-size Central City	Low Percent Minority	10
Urban Fringe of Mid-size Central City	Medium Percent Minority	10
Urban Fringe of Mid-size Central City	High Percent Minority	9
Large/Small Town	Low Percent Minority	9
Large/Small Town	Medium Percent Minority	9
Large/Small Town	High Percent Minority	9
Rural	Low Percent Minority	16
Rural	Medium Percent Minority	14
		112
ARIZONA (Small School Cluster Type 2 - Geographic)		
Large Central City	Low Percent Minority	9
Large Central City	Medium Percent Minority	8
Large Central City	High Percent Minority	9
Mid-size Central City	Low Percent Minority	10
Mid-size Central City	Medium Percent Minority	10
Mid-size Central City	High Percent Minority	9
Urban Fringe of Large Central City	Low Percent Minority	6
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	7
Urban Fringe of Large/Mid-size Central City	High Percent Minority	6
Large/Small Town and Rural	Low Percent Minority	13
Large/Small Town and Rural	Medium Percent Minority	13
Large/Small Town and Rural	High Percent Minority	10
		110
ARKANSAS (Small School Cluster Type 2 - Geographic)		
Mid-size Central City+Urban Fringe	Low Percent Minority	10
Mid-size Central City+Urban Fringe	Medium Percent Minority	10
Mid-size Central City+Urban Fringe	High Percent Minority	10
L/Small Town	Low Percent Minority	15
L/Small Town	Medium Percent Minority	14
L/Small Town	High Percent Minority	15
Rural	Low Percent Minority	19
Rural	Medium Percent Minority	15
Rural	High Percent Minority	16
		124

Table 3-2 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	In-scope Schools in Strata
CALIFORNIA (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	13
Large/Mid-size Central City	Medium Percent Minority	12
Large/Mid-size Central City	High Percent Minority	13
Urban Fringe of Large Central City	Low Percent Minority	10
Urban Fringe of Large Central City	Medium Percent Minority	11
Urban Fringe of Large Central City	High Percent Minority	11
Urban Fringe of Mid-size Central City	Low Percent Minority	5
Urban Fringe of Mid-size Central City	Medium Percent Minority	5
Urban Fringe of Mid-size Central City	High Percent Minority	4
Large/Small Town and Rural	Low Percent Minority	13
Large/Small Town and Rural	Medium Percent Minority	9
Large/Small Town and Rural	High Percent Minority	7
		113
COLORADO (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	12
Large/Mid-size Central City	Medium Percent Minority	11
Large/Mid-size Central City	High Percent Minority	11
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	14
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	14
Urban Fringe of Large/Mid-size Central City	High Percent Minority	14
Large/Small Town	Low Percent Minority	7
Large/Small Town	Medium Percent Minority	7
Large/Small Town	High Percent Minority	6
Rural	Low Percent Minority	11
Rural	Medium Percent Minority	9
Rural	High Percent Minority	11
		127
CONNECTICUT (Small School Cluster Type 2 - Geographic)		
Large Central City	Low Black/Low Hispanic	5
Large Central City	Low Black/High Hispanic	4
Large Central City	High Black/Low Hispanic	4
Large Central City	High Black/High Hispanic	4
Mid-size Central City	Low Percent Minority	7
Mid-size Central City	Medium Percent Minority	7
Mid-size Central City	High Percent Minority	7
Urban Fringe of Large/Mid-size Central City	None	34
Large/Small Town and Rural	None	39
		111

Table 3-2 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	In-scope Schools in Strata
DELAWARE (Small School Cluster Type 1 - None)		
Large/Mid-size Central City	Low Percent Minority	3
Large/Mid-size Central City	Medium Percent Minority	4
Large/Mid-size Central City	High Percent Minority	5
Urban Fringe of Mid-size Central City	Low Percent Minority	1
Urban Fringe of Mid-size Central City	Medium Percent Minority	2
Urban Fringe of Mid-size Central City	High Percent Minority	3
Small Town	Low Percent Minority	2
Small Town	Medium Percent Minority	3
Small Town	High Percent Minority	1
Rural	Low Percent Minority	7
Rural	Medium Percent Minority	8
Rural	High Percent Minority	8
		47
DISTRICT OF COLUMBIA (Small School Cluster Type 1 - None)		
Large Central City	Medium Percent Minority	44
Large Central City	High Percent Minority	69
		113
FLORIDA (Small School Cluster Type 2 - Geographic)		
Large Central City	Low Black/Low Hispanic	4
Large Central City	Low Black/High Hispanic	4
Large Central City	High Black/Low Hispanic	4
Large Central City	High Black/High Hispanic	4
Mid-size Central City	Low Percent Minority	6
Mid-size Central City	Medium Percent Minority	7
Mid-size Central City	High Percent Minority	7
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	16
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	16
Urban Fringe of Large/Mid-size Central City	High Percent Minority	15
Large/Small Town and Rural	Low Percent Minority	8
Large/Small Town and Rural	Medium Percent Minority	8
Large/Small Town and Rural	High Percent Minority	7
		106
GEORGIA (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	8
Large/Mid-size Central City	Medium Percent Minority	8
Large/Mid-size Central City	High Percent Minority	8
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	10
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	11
Urban Fringe of Large/Mid-size Central City	High Percent Minority	10
Large/Small Town	Low Percent Minority	11
Large/Small Town	Medium Percent Minority	11
Large/Small Town	High Percent Minority	10
Rural	Low Percent Minority	7
Rural	Medium Percent Minority	6
Rural	High Percent Minority	7
		107

Table 3-2 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	In-scope Schools in Strata
GUAM (Small School Cluster Type 1 - None)		
Rural	None	21
HAWAII (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	None	33
Urban Fringe of Mid-size Central City	None	51
Large/Small Town and Rural	None	23
		107
IDAHO (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City and Urban Fringe	None	22
Large Town	None	19
Small Town	None	35
Rural	None	39
Small Schools		
None	None	14
		129
INDIANA (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	12
Large/Mid-size Central City	Medium Percent Minority	11
Large/Mid-size Central City	High Percent Minority	10
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	13
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	11
Rural	None	26
Large/Small Town	None	33
		116
IOWA (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City and Urban Fringe	None	38
Large/Small Town	None	40
Rural	None	47
Small Schools		
None	None	14
		139
KENTUCKY (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	6
Mid-size Central City	Medium Percent Minority	7
Mid-size Central City	High Percent Minority	6
Urban Fringe of Mid-size Central City	Low Percent Minority	9
Urban Fringe of Mid-size Central City	Medium Percent Minority	8
Rural	None	51
Large/Small Town	None	36
		123

Table 3-2 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	In-scope Schools in Strata
LOUISIANA (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	11
Large/Mid-size Central City	Medium Percent Minority	11
Large/Mid-size Central City	High Percent Minority	12
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	6
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	7
Urban Fringe of Large/Mid-size Central City	High Percent Minority	6
Large/Small Town	Low Percent Minority	11
Large/Small Town	Medium Percent Minority	11
Large/Small Town	High Percent Minority	11
Rural	Low Percent Minority	8
Rural	Medium Percent Minority	11
Rural	High Percent Minority	9
		114
MAINE (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City and Urban Fringe	None	21
Small Town	None	58
Rural	None	45
Small Schools		
None	None	39
		163
MARYLAND (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	7
Large/Mid-size Central City	Medium Percent Minority	6
Large/Mid-size Central City	High Percent Minority	7
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	21
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	22
Urban Fringe of Large/Mid-size Central City	High Percent Minority	21
Large/Small Town and Rural	Low Percent Minority	14
Large/Small Town and Rural	Medium Percent Minority	12
		110
MASSACHUSETTS (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	14
Large/Mid-size Central City	Medium Percent Minority	13
Large/Mid-size Central City	High Percent Minority	13
Urban Fringe of Large/Mid-size Central City	None	40
Large/Small Town and Rural	None	40
		120
MICHIGAN (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	9
Large/Mid-size Central City	Medium Percent Minority	8
Large/Mid-size Central City	High Percent Minority	9
Urban Fringe of Large/Mid-size Central City	None	38
Large/Small Town	None	30
Rural	None	20
		114

Table 3-2 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	In-scope Schools in Strata
MINNESOTA (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Medium Percent Minority	5
Large/Mid-size Central City	Low Percent Minority	7
Urban Fringe of Large/Mid-size Central City	None	36
Large/Small Town	None	26
Rural	None	41
		115
MISSISSIPPI (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	4
Mid-size Central City	Medium Percent Minority	5
Mid-size Central City	High Percent Minority	4
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	3
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	3
Urban Fringe of Large/Mid-size Central City	High Percent Minority	3
Large/Small Town	Low Percent Minority	15
Large/Small Town	Medium Percent Minority	15
Large/Small Town	High Percent Minority	15
Rural	Low Percent Minority	13
Rural	Medium Percent Minority	13
Rural	High Percent Minority	15
		108
MISSOURI (Small School Cluster Type 3 - Stratified)		
Large Schools		
Large/Mid-size Central City	Low Percent Minority	5
Large/Mid-size Central City	Medium Percent Minority	6
Large/Mid-size Central City	High Percent Minority	3
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	13
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	13
Urban Fringe of Large/Mid-size Central City	High Percent Minority	13
Large/Small Town	None	24
Rural	None	33
Small Schools		
None	None	13
		123
NEBRASKA (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City and Urban Fringe	None	43
Large/Small Town	None	37
Rural	None	41
Small Schools		
None	None	66
		187

Table 3-2 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	In-scope Schools in Strata
NEW HAMPSHIRE (Small School Cluster Type 3 - Stratified)		
Large Schools Mid-size Central City and Urban Fringe	None	26
Large/Small Town	None	57
Rural	None	29
Small Schools None	None	24
		136
NEW JERSEY (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Black/Low Hispanic	6
Large/Mid-size Central City	Low Black/High Hispanic	5
Large/Mid-size Central City	High Black/Low Hispanic	5
Large/Mid-size Central City	High Black/High Hispanic	5
Urban Fringe of Large Central City	Low Percent Minority	28
Urban Fringe of Large Central City	Medium Percent Minority	17
Urban Fringe of Mid-size Central City	None	25
Large/Small Town and Rural	None	28
		119
NEW MEXICO (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City and Urban Fringe	Low Percent Minority	14
Mid-size Central City and Urban Fringe	Medium Percent Minority	14
Mid-size Central City and Urban Fringe	High Percent Minority	14
Large Town	Low Percent Minority	5
Large Town	Medium Percent Minority	5
Large Town	High Percent Minority	6
Small Town	Low Percent Minority	10
Small Town	Medium Percent Minority	10
Small Town	High Percent Minority	11
Rural	Low Percent Minority	5
Rural	Medium Percent Minority	7
Rural	High Percent Minority	8
Small Schools		
None	None	11
		120
NEW YORK (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	High Black/High Hispanic	11
Large/Mid-size Central City	Low Black/Low Hispanic	12
Large/Mid-size Central City	Low Black/High Hispanic	12
Large/Mid-size Central City	High Black/Low Hispanic	12
Urban Fringe of Large Central City	None	13
Urban Fringe of Mid-size Central City	None	18
Large/Small Town and Rural	None	32
		110

Table 3-2 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	In-scope Schools in Strata
NORTH CAROLINA (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	10
Mid-size Central City	Medium Percent Minority	9
Mid-size Central City	High Percent Minority	10
Urban Fringe of Mid-size Central City	Low Percent Minority	4
Urban Fringe of Mid-size Central City	Medium Percent Minority	5
Urban Fringe of Mid-size Central City	High Percent Minority	4
Large/Small Town	Low Percent Minority	11
Large/Small Town	Medium Percent Minority	11
Large/Small Town	High Percent Minority	10
Rural	Low Percent Minority	17
Rural	Medium Percent Minority	12
Rural	High Percent Minority	12
		115
NORTH DAKOTA (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City and Urban Fringe	None	36
Large/Small Town	None	31
Rural	None	51
Small Schools		
None	None	42
		160
OHIO (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	11
Large/Mid-size Central City	Medium Percent Minority	10
Large/Mid-size Central City	High Percent Minority	11
Urban Fringe of Large/Mid-size Central City	None	32
Large/Small Town	None	24
Rural	None	29
		117
OKLAHOMA (Small School Cluster Type 3 - Stratified)		
Large Schools		
Large/Mid-size Central City	Low Percent Minority	16
Large/Mid-size Central City	Medium Percent Minority	17
Urban Fringe of Large/Mid-size Central City	None	14
Large/Small Town	None	37
Rural	None	34
Small Schools		
None	None	23
		141

Table 3-2 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	In-scope Schools in Strata
PENNSYLVANIA (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	9
Large/Mid-size Central City	Medium Percent Minority	9
Large/Mid-size Central City	High Percent Minority	8
Urban Fringe of Large/Mid-size Central City	None	33
Large/Small Town	None	36
Rural	None	23
		118
RHODE ISLAND (Small School Cluster Type 2 - Geographic)		
Large Central City	Low Percent Minority	8
Large Central City	Medium Percent Minority	6
Large Central City	High Percent Minority	5
Mid-size Central City	None	9
Urban Fringe of Large/Mid-size Central City	None	55
Large/Small Town and Rural	None	27
		110
SOUTH CAROLINA (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	6
Mid-size Central City	Medium Percent Minority	5
Mid-size Central City	High Percent Minority	6
Urban Fringe of Mid-size Central City	Low Percent Minority	10
Urban Fringe of Mid-size Central City	Medium Percent Minority	10
Urban Fringe of Mid-size Central City	High Percent Minority	10
Small Town	Low Percent Minority	13
Small Town	Medium Percent Minority	12
Small Town	High Percent Minority	12
Rural	Low Percent Minority	9
Rural	Medium Percent Minority	9
Rural	High Percent Minority	9
		111
TENNESSEE (Small School Cluster Type 2 - Geographic)		
Large/Mid-size Central City	Low Percent Minority	13
Large/Mid-size Central City	Medium Percent Minority	13
Large/Mid-size Central City	High Percent Minority	12
Urban Fringe of Large/Mid-size Central City	None	19
Large/Small Town	None	31
Rural	None	32
		120

Table 3-2 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	In-scope Schools in Strata
TEXAS (Small School Cluster Type 2 - Geographic)		
Large Central City	Low Hispanic/Low Black	7
Large Central City	Low Hispanic/High Black	6
Large Central City	High Hispanic/Low Black	7
Large Central City	High Hispanic/High Black	6
Mid-size Central City	Low Percent Minority	7
Mid-size Central City	Medium Percent Minority	8
Mid-size Central City	High Percent Minority	9
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	7
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	5
Urban Fringe of Large/Mid-size Central City	High Percent Minority	5
Large/Small Town and Rural	Low Percent Minority	15
Large/Small Town and Rural	Medium Percent Minority	14
Large/Small Town and Rural	High Percent Minority	15
		111
UTAH (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	None	26
Urban Fringe of Mid-size Central City	None	46
Large/Small Town	None	15
Rural	None	24
		111
VIRGINIA (Small School Cluster Type 2 - Geographic)		
Mid-size Central City	Low Percent Minority	13
Mid-size Central City	Medium Percent Minority	11
Mid-size Central City	High Percent Minority	12
Urban Fringe of Large/Mid-size Central City	Low Percent Minority	11
Urban Fringe of Large/Mid-size Central City	Medium Percent Minority	10
Urban Fringe of Large/Mid-size Central City	High Percent Minority	9
Large/Small Town	Medium Percent Minority	5
Large/Small Town	High Percent Minority	6
Large/Small Town	Low Percent Minority	5
Rural	Low Percent Minority	9
Rural	Medium Percent Minority	11
Rural	High Percent Minority	8
		110
VIRGIN ISLANDS (Small School Cluster Type 1 - None)		
	Low Percent Minority	10
	Medium Percent Minority	6
	High Percent Minority	8
		24

Table 3-2 (continued)
Distribution of the Selected Schools by Sampling Strata, Grade 4

Urbanization	Minority	In-scope Schools in Strata
WEST VIRGINIA (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City	None	18
Urban Fringe of Mid-size Central City	None	19
Large/Small Town	None	35
Rural	None	65
Small Schools		
None	None	19
		156
WISCONSIN (Small School Cluster Type 3 - Stratified)		
Large Schools		
Large/Mid-size Central City	Low Percent Minority	17
Large/Mid-size Central City	Medium Percent Minority	17
Urban Fringe of Large/Mid-size Central City	None	20
Large/Small Town	None	31
Rural	None	32
Small Schools		
None	None	12
		129
WYOMING (Small School Cluster Type 3 - Stratified)		
Large Schools		
Mid-size Central City	None	18
Urban Fringe of Mid-size Central City	None	15
Large/Small Town	None	62
Rural	None	25
Small Schools		
None	None	60
		180

5) *Large Town:* a place not within an MSA, but with a population greater than or equal to 25,000 and defined as urban by the U.S. Bureau of Census.

6) *Small Town:* a place not within an MSA, but with a population less than 25,000 and defined as urban by U.S. Bureau of Census.

7) *Rural:* a place with a population of less than 2,500 and defined as rural by the U.S. Bureau of the Census.

The urbanization strata were created by collapsing type of locale categories. The nature of the collapsing varied across states and grades. Each urbanization stratum included a minimum of 10 percent of eligible students in the participating state. Table 3-2 provides the urbanization categories (created by collapsing type of locale) used within each state.

3.4.3 Minority Classification

The third stage of stratification was minority enrollment. Minority enrollment strata were formed within urbanization strata, based on the percentages of Black and Hispanic students. The three cases that occur are described in the following paragraphs.

Case 1: Urbanization strata with less than 10 percent Black students and 7 percent Hispanic students were not stratified by minority enrollment.

Case 2: Urbanization strata with more than 10 percent Black students or 7 percent Hispanic students, but not more than 20 percent of each, were stratified by ordering percent minority enrollment within the urbanization classes and dividing the schools into three groups with about equal numbers of students per minority group.

Case 3: In urbanization strata with more than 20 percent of both Black and Hispanic students, minority strata were formed with the objective of providing equal strata with emphasis on the minority group (Black or Hispanic) with the higher concentration. The stratification was performed as follows. The minority group with the higher percentage gave the primary stratification variable; the remaining group gave the secondary stratification variable. Within urbanization class, the schools were sorted based on the primary stratification variable and divided into two groups of schools containing approximately equal numbers of students. Within each of these two groups, the schools were sorted by the secondary stratification variable and subdivided into two subgroups of schools containing approximately equal numbers of students. As a result, within urbanization strata there were four minority groups, low Black/low Hispanic, low Black/high Hispanic, high Black/low Hispanic, and high Black/high Hispanic.

The minority groups (with almost equal sizes) were formed solely for the purpose of creating efficient stratification design at this stage of sampling. These classifications are not directly used in analysis and reporting of the data, but will act to reduce sampling errors for achievement-level estimates. Table 3-2 provides information on minority stratification for the participating states.

3.4.4 Median Household Income

The median household income variable was not used as a prime stratification variable because the available income data were not up to date (i.e., they were based on the 1980 Census). Instead, median household income was used as a sorting variable at the final stage of stratification. Prior to the selection of the school samples, the schools were sorted by urbanization, then by minority classes within urbanization in a serpentine order, in which the sort alternated between descending and ascending order within each group. This meant that adjacent schools on the list were generally similar with regard to either urbanization or minority enrollment, and often to both. Within minority class, the schools were sorted, in serpentine order, by the median household income. This final stage of sorting resulted in implicit stratification of median income. The data on median household income, which were obtained from Donnelly Marketing Information Services, were related to the ZIP code area in which the school was located. The data are derived from the 1980 census, but expressed in 1985 dollars.

3.4.5 Schools With Fewer Than 20 Students

Schools with fewer than 20 students were combined with other schools to form a sampling unit of at least 20 students. The two methods used to combine small schools are referred to as geographic and stratified grouping.

Geographic Grouping. If the number of small schools in the state was less than 20 percent, and the number of students in these small schools accounted for less than 1 percent of the total state grade enrollment, then each school was combined with a school close by geographically until the cluster contained at least 20 students.

Cluster level values for enrollment, urbanization, minority, income variables, and selection probabilities was equal to the corresponding values of the school in the cluster with largest enrollment.

Stratified Grouping. In states with a larger number of small schools (Cluster Type 3 states), schools were stratified into two groups. One group contained schools with fewer than 20 students, the other group contained schools with 20 or more students. The schools in the first group were clustered in the following manner. The schools were ordered from smallest to largest, then the largest school was matched with the smallest school. If this cluster contained 20 or more students, it was complete. If the total cluster enrollment was 19 or smaller, the next smallest school was added. This continued until the sum of the enrollment was at least 20. The next cluster was formed with the next largest and smallest school in the same manner. If, after forming all the clusters, there remained a cluster with fewer than 20 students, it was combined with the previous cluster.

The enrollment value assigned to a cluster was equal to the sum of enrollments of the schools in that cluster. The minority value assigned to the cluster was equal to the weighted average of the proportion of minority for schools in the cluster where the weight was the fourth-grade enrollment. The cluster level income value was the median income of the school with the largest enrollment. No urbanization value was desired for clusters of schools. Also, no selection probability was derived for clusters of schools since they were selected with equal probability.

Table 3-2 shows the type of stratification used for small schools within the participating states.

3.5 SCHOOL SAMPLE SELECTION FOR THE 1992 TRIAL STATE ASSESSMENT

3.5.1 Control of Overlap of School Samples for National Educational Studies

The issue of school sample overlap has been relevant in all rounds of NAEP in recent years, but no more so than in 1992. NAEP collected data nationally from a number of distinct samples at all three age classes, while state assessments were conducted at grades 4 and 8. At the same time, the U.S. Department of Education conducted the first phase followup to *Prospects: The National Longitudinal Study of Chapter I Children* (Abt Associates, 1991), for which a sample of districts was selected prior to the 1992 Trial State Assessment sample selection. This study involved substantial student assessment at grades 4 and 8.

To avoid undue burden on individual schools, NAEP developed a policy for 1992 of avoiding overlap of school samples from different studies for the same age class. This was to be achieved without unduly distorting the resulting samples by introducing bias or substantial variance. Thus, at grade 4 for example, the school samples for the national samples, the state samples, and the Prospects samples were selected to contain different schools, to the extent feasible. The procedure used was an extension of the method proposed by Keyfitz (1951). The general approach is given in the remainder of this section. Besides generally controlling overlap within grade, distinct schools were selected for the fourth- and eighth-grade state assessment samples within a state to the extent feasible.

Consider as an example the selection of samples for the Trial State Assessment fourth-grade sample. At the time of drawing the NAEP samples, the identities of the Prospects sample schools were not known. Since the selected districts, and district selection probabilities for all districts, were known, this information was used to control sample overlap. For each school in the frame for the national and state NAEP samples, there was a flag, C, indicating whether ($C=1$) or not ($C=0$) the district containing the school was included in the Prospects sample, and a Prospects district selection probability, $P_c = P(C=1)$.

In controlling overlap between NAEP state and national sample school selections, national school selection probabilities that were conditional on the selection of national sample PSUs (i.e., the school-within-PSU selection probabilities) were used. This meant that in selecting the state samples, in those states where there was no PSU selection for the national samples no adjustments were needed to account for the selection of national NAEP samples (which might have selected schools within that state but, in fact, did not). This procedure of conditioning on the selection of PSUs also recognizes the impact of the heavy within-PSU sampling in noncertainty PSUs in some states, even though the unconditional probabilities of selection for such schools in the national samples were quite low. In other words, conditioning on the national PSU sample reduces the variance of the state samples, although it leads to a greater degree of sample overlap than if unconditional national selection probabilities had been used in the procedure for controlling overlap between state and national samples.

Let N = 1 if the school is selected in the national sample; let N = 0 otherwise. Let $P_N = P(N = 1)$. Thus, $P_N = 0$ for schools not located within a selected national sample PSU. Let π_s denote the probability that a school is to be selected for the state fourth-grade sample. Schools to be included with certainty in the state sample ($\pi_s = 1$) are not subject to overlap control, as such schools are self-representing in the state sample. Excluding such schools on a random basis would add undue variance to the state estimates.

Where possible, schools in districts selected for the Prospects study were excluded provided that the Prospects' district selection probability, P_c, fell below a constant, k_s, that varied from state to state. In small states, where it is important to include all eligible schools in the state sample, k_s was set to zero. The variable C indicates whether (C=1) or not (C=0) the district was included in the Prospects sample. For actually drawing the state samples, a conditional probability of selection, π_s^*, was derived as follows for each school in the frame having grade 4 enrollment but no grade 8 enrollment:

$$
\begin{array}{ll}
\pi_s^* = 1 & \text{if } \pi_s = 1 \\
\pi_s / (1 - \phi_N) & \text{if } \pi_s < 1,\ P_C > k_s, \text{ and } N = 0 \\
\dfrac{\pi_s \, (P_N - \phi_N)}{P_N \, (1 - \phi_N)} & \text{if } \pi_s < 1,\ P_C > k_s, \text{ and } N = 1 \\
\pi_s / (1 - \nu_{NC}) & \text{if } \pi_s < 1,\ P_C \le k_s, \text{ and } (C = 0 \text{ and } N = 0) \\
\dfrac{\pi_s (P_N + P_C - \nu_{NC})}{(P_N + P_C)(1 - \nu_{NC})} & \text{if } \pi_s < 1,\ P_C \le k_s, \text{ and } (C = 1 \text{ or } N = 1)
\end{array}
$$

where $\phi_N = \min(P_N, 1 - \pi_s)$ and $\nu_{NC} = \min(P_N, P_C, 1 - \pi_s)$. The values of π_s^* are conditional on the selection of districts for the Prospects sample and PSUs for the national NAEP samples. For schools having enrollment at both fourth and eighth grades, an analogous (but more complicated) formula was used to derive π_s^* in a way that also minimized the overlap of the fourth-grade school sample with the eighth-grade school sample within the state.

This procedure in general gave state NAEP conditional selection probabilities that are smaller than the unconditional selection probabilities for schools located in Prospects selected districts, and for schools selected for the national sample. The relative chance of selection in the state sample for a school selected in either of these other two samples, compared to its chance of selection in the state sample if not selected for either of the other samples, is $(P_N - \phi_N)/P_N$ if $P_C > k_s$ and $(P_N + P_C - \nu_{NC})/(P_N + P_C)$ if $P_C \le k_s$. If P_N, P_C, and π_s are all relatively small, then $P_N + P_C - \nu_{NC} = 0$, so that there was no chance of selecting the school for the state sample if it is in the national sample or in a Prospects district selection. The probability that a school was selected in the state sample, conditional on the national PSU sample but unconditional on the national school sample selection within PSUs and the selection of districts for the Prospects sample, is given by π_s, as desired. This follows from the above formulation of π_s^* and the fact that P(C=1 or N=1) equals $P_N + P_C$ when $P_C \le k_s$ since there is no overlap of NAEP national sample selected schools and Prospects selected districts in this case. The quantity π_s is used as the basis for weighting the schools, and hence students, in the state samples.

To illustrate the implementation of these expressions for drawing the state sample, suppose that $P_C > k_s$ (or $P_C=0$) so that one is concerned only with controlling overlap with the national sample. Suppose that $\pi_s = 0.3$ and $P_N = 0.25$. Then $\phi_N = P_N = 0.25$, and $\pi_s^* = 0.4$ if the school is not selected for the national sample. Thus in this case the school is selected with probability 0.4. Since $\phi_N = P_N$, $\pi_s^* = 0$ if the school is selected for the national sample. Thus there is no chance that this school will be selected for both the national and state samples. Integrating over the national sampling process gives the required unconditional state selection probability of 0.3 (= 0.4 * (1 - 0.25) + 0 * 0.25).

3.5.2 Selection of Schools in Small States (Cluster Type 1 States)

For states with small numbers of schools, and no or very few small schools, all schools were included in the sample with certainty. All the eligible fourth-grade schools in the District of Columbia, Delaware, Guam, and the Virgin Islands were taken into the sample with certainty.

3.5.3 States with Geographic Clustering of Small Schools (Cluster Type 2 States)

Clusters were sorted by urbanization, minority strata (which varied by state and urbanization level), and median income. A systematic sample of clusters was then selected for each state with probability proportionate to size, where size was equal to the estimated grade enrollment within the school, so as to achieve the desired student sample size of 6,300. The fourth-grade sample selected two sessions from larger schools (those with 20 or more students), one for reading and one for mathematics assessments.

Following the selection of clusters, there was some thinning of small schools. The purpose of thinning was to give students in small schools (enrollment of fewer than 20) approximately the same chance of selection as those from larger schools, and to control the sample size of schools to be close to the desired number. All small schools in a cluster were retained in the sample with probability P_s/P_c where P_s was the probability of selection of the small school and P_c was the probability of selection of the cluster.

Table 3-3 shows the distribution of selected schools in the participating states.

3.5.4 States with Stratification of Small Schools (Cluster Type 3 States)

As described above, clusters were sorted by urbanization, minority strata (which varied by state and urbanization level), and median income within the two size clusters. Small school clusters were selected systematically with equal probability, and large schools were sampled systematically with probability proportionate to size, so as to achieve the desired student sample size of 6,300 for the fourth grade.

Table 3-3
Distribution of Sample Sizes by School Size, with Corresponding Overlap Between Grades

State	Number of Small* Schools Sampled for...		Number of Other Schools Sampled for...	
	4th Only	**4th & 8th**	**4th Only**	**4th & 8th**
Alabama	2	0	113	0
Arizona	6	0	106	0
Arkansas	7	0	119	0
California	8	0	108	0
Colorado	15	0	114	0
Connecticut	1	0	114	0
Delaware	0	2	50	2
District of Columbia	3	0	108	7
Florida	1	0	106	0
Georgia	0	0	109	0
Guam	0	0	21	0
Hawaii	1	0	94	12
Idaho	15	0	115	0
Indiana	2	0	116	0
Iowa	14	0	129	0
Kentucky	4	0	122	0
Louisiana	7	0	111	0
Maine	40	0	122	3
Maryland	2	0	109	0
Massachusetts	1	0	122	0
Michigan	2	0	114	0
Minnesota	4	0	116	0
Mississippi	1	0	110	0
Missouri	14	0	116	0
Nebraska	79	0	121	0
New Hampshire	25	0	115	0
New Jersey	4	0	119	0
New Mexico	12	0	110	0
New York	1	0	109	0
North Carolina	4	0	113	0
North Dakota	46	0	118	1
Ohio	2	0	116	0
Oklahoma	26	0	118	0
Pennsylvania	0	0	118	0
Rhode Island	0	1	108	6
South Carolina	2	0	111	0
Tennessee	7	0	115	0
Texas	5	0	109	0
Utah	6	0	106	0
Virginia	3	0	111	0
Virgin Islands	1	0	22	1
West Virginia	26	0	140	0
Wisconsin	14	0	120	0
Wyoming	61	12	120	2

---

*Small school denotes a school with fewer than 20 fourth-grade students.

Similar to Cluster Type 2 states, each selected fourth-grade school was chosen for one reading and one mathematics session except for schools with fourth-grade enrollment of fewer than 20, which were assigned only a single session.

Table 3-3 shows the distribution of all selected schools in the participating states, including some schools that were found to have no eligible students after sample selection (out-of-scope).

3.5.5 Overlap of School Samples

As stated in section 3.5.1, the sample design for fourth-grade schools minimized, to the extent feasible, the chances of selecting fourth-grade schools in the 1992 national NAEP and the Prospects survey. Furthermore, the fourth-grade state samples were selected such that the number of schools in each state selected for both fourth- and eighth-grade samples were minimized to the extent feasible.

Table 3-3 shows the overlap of fourth- and eighth-grade schools in participating states. Only three schools were selected for both state and national fourth-grade samples—one each in Arkansas, New York, and Wyoming.

3.5.6 New School Selection

A district-level file was constructed from the aggregate of the fourth- and eighth-grade school frame. The file was divided into a small districts file, consisting of those districts in which there were at most two schools on the aggregate frame but no more than one fourth- and one eighth-grade school. The remainder of districts were denoted as "large" districts.

A sample of "large" districts was drawn in each state. All districts were selected in Delaware, the District of Columbia, Guam, Hawaii, and the Virgin Islands. The remainder of the states in the file of "large" districts (eligible for sampling) was divided into two files within each state. Two districts were selected with equal probability among the smaller districts with combined enrollment of about 20 percent of the state enrollment. From the rest of the file, eight districts per state were selected with probability proportional to enrollment. The 10 selected districts were then sent a listing of all their schools that appeared on the QED sampling frame, and were asked to provide information about the new schools not included in the QED frame. These listings, provided by selected districts, were used as sampling frames for selection of new schools.

The eligibility of a school was determined based on the grade span. A school was classified as "new " if the changes of grade span were such that the school status changed from ineligible to eligible. The average grade enrollment for these schools was set to the average grade enrollment before the grade span change. The schools found eligible for sampling due to the grade span change were added to the corresponding grade frame.

Each fourth-grade school was assigned the measure of size:

$$\begin{cases} 60 & \text{if enrollment} \leq 70 \\ \textit{enrollment} & \text{if enrollment} > 70 \end{cases}$$

The probability of selecting a school was $\min\left[\frac{\text{sampling rate} * \text{measure of size}}{\text{P(district)}}, 1\right]$,

where P(district) was the probability of selection of a district and the sampling rate was the rate used for the particular state in the selection of the original sample of schools.

In each state, the sampling rate used for the main sample of fourth-grade schools was used to select the new schools. Additionally, all new eligible schools coming from "small" districts (those with at most one grade 4 and one grade 8 school) that had a school selected in the regular sample for the fourth grade were included in the sample and treated as belonging to the same cluster as the original selection from that district.

Table 3-4 shows the number of new schools coming from the "large" and "small" districts for the fourth-grade samples.

3.5.7 Assigning Subject Session Types

In the interest of sampling efficiency it was desirable that each of the two subjects assessed at grade 4, reading and mathematics, be administered in as large a subset of the sampled schools as possible. On the other hand it was unreasonable to expect very small schools to conduct two different sessions with half of the eligible students in each. To satisfy these two requirements the following procedure was used.

If, according to the information on the frame, the school had an enrollment of 21 or more grade 4 students, the school was assigned initially to conduct both mathematics and reading sessions, with half of the selected students being assigned to a mathematics session, and half to a reading session (see section 3.6 for a description of the student sampling process). This varied only in Guam, where all students took both assessment types.

If, according to the frame data, the school enrollment was 20 or fewer, the school was randomly assigned to conduct either a mathematics or a reading session. The assignment was systematic, based on the ordering of the clusters for sample selection, with random ordering of selected schools within clusters.

If a school had two session types assigned initially, but was found at the time of drawing the student samples to have fewer than 21 eligible students, the school was randomly assigned to conduct only one of the two session types, with each type being chosen with probability 0.5. This assignment was independent from school to school. Thus a school was to conduct a single

Table 3-4
Distribution of New Schools Coming from "Large" and "Small" Districts in the Fourth-grade Sample

State	Number of New Schools	
	"Large" Districts	"Small" Districts
Alabama	-	-
Arizona	-	-
Arkansas	1	-
California	2	1
Colorado	2	-
Connecticut	-	-
Delaware	3	-
District of Columbia	1	-
Florida	5	-
Georgia	1	-
Guam	-	-
Hawaii	1	-
Idaho	-	-
Indiana	1	-
Iowa	-	-
Kentucky	2	-
Louisiana	-	-
Maine	-	-
Maryland	2	-
Massachusetts	-	1
Michigan	1	-
Minnesota	1	-
Mississippi	1	-
Missouri	1	-
Nebraska	-	-
New Hampshire	-	-
New Jersey	1	-
New Mexico	-	-
New York	-	-
North Carolina	4	-
North Dakota	1	-
Ohio	5	-
Oklahoma	-	-
Pennsylvania	1	-
Rhode Island	-	-
South Carolina	1	-
Tennessee	2	-
Texas	-	-
Utah	-	1
Virginia	5	-
Virgin Islands	-	-
West Virginia	1	-
Wisconsin	1	-
Wyoming	-	-

session type if either its frame or its actual enrollment for grade 4 was 20 or fewer; a school was to conduct both session types if both its frame and actual enrollments exceeded 20.

3.5.8 Designating Monitor Status

Within each state, random equivalent half samples of schools were assigned to be monitored or unmonitored. The details of the implementation of the monitoring process in the field are given in Chapter 4. The purpose of monitoring a random half of the schools was to ensure that the procedures were being followed throughout each state by the school and district personnel administering the assessments, and to provide data adequate for assessing whether there was a significant difference in assessment results between monitored and unmonitored schools within each state.

The following procedure was used to determine the sample of schools to be monitored. The initially selected clusters were sorted in the order in which they were systematically selected (see sections 3.5.2 to 3.5.4). New schools from "large" districts added to the sample (see section 3.5.6) were treated as single school clusters, and were added to the end of the list in random order. The sorted clusters were then paired, and one member of each pair was assigned at random, with probability 0.5, to be monitored. The assignment was independent across pairs. If there was an odd number of clusters, the last cluster was assigned, with 0.5 probability, to be monitored.

If a cluster was designated to be monitored, all selected schools within the cluster (after thinning of small schools from multiple school clusters in Cluster Type 2 states; see section 3.5.3) were assigned to be monitored. For the grade 4 samples, this procedure, in combination with the procedure for assigning schools to subjects (see section 3.5.7), ensured that for every pair of clusters for each subject at least one school would be monitored and at least one would not.

In the territories of Guam and the Virgin Islands, there were few schools in each sample, and large samples of students (that is, all of the students enrolled) were drawn from each school. In these jurisdictions the monitoring assignment was done at the level of the physical assessment session, rather than at the cluster level. After establishing in each school the number of sessions to be conducted, alternate sessions were designated to be monitored, with the first session assigned at random. Thus all schools contained some monitored and some unmonitored sessions.

3.5.9 School Substitution and Participation

A substitute school was selected for each selected school containing eligible students, for which school nonparticipation was established by the state coordinator as of November 1, 1991. The process of selecting a substitute for a school involved identifying the most similar school in terms of the following characteristics: urbanization, percent of Black enrollment, percent of Hispanic enrollment, fourth-grade enrollment, and median income. To identify candidates for substitution, a set of schools was found that provided reasonable matches with regard to fourth-grade enrollment, and percent of Black and Hispanic enrollment. From among this set a match was selected, considering all five characteristics. Schools in the National Assessment sample and those in the Prospects study were avoided, where possible, in the selection of substitutes. Furthermore, the substitute was selected from the same district, where possible, to avoid placing the burden of replacing a refusing school from one district on another district. This was often not possible, however, as in the majority of cases the decision not to participate was made at the district level.

In the cases where no suitable substitute could be found among those schools not sampled (most often because all or most schools had been included in the original sample), a school already in the sample was selected to conduct a double session, of which one session served as a substitute for students in the refusing school. The same criteria were applied in selecting the schools that conducted double sessions; that is, a reasonable match was found based on grade enrollment, percent of Black and Hispanic enrollment, median income, and urbanization.

Table 3-5 includes information about the number of substitutes provided in each state. Of the 44 states participating, 27 were provided with at least one substitute. Among states receiving no substitutes, the majority had 100 percent participation from the original sample. In a few cases, however, refusals did occur after the November 1 deadline. The number of substitutes provided to a state ranged from 0 to 59 in the fourth grade sample. A total of 591 substitutes were selected for the fourth-grade sample, 23 of which were double session substitutes. Some states did not attempt to solicit participation from the substitute schools provided, as they considered the timing too late to seek cooperation from schools not previously notified about the assessment. In quite a few cases the originally selected school agreed to cooperate after a substitute was selected and had agreed to participate (in which case the substitute school data were discarded).

Table 3-6 shows the number of schools in the fourth-grade reading samples, together with school response rates observed within participating states. The table also shows the number of substitutes in each state that were associated with a nonparticipating original school selection, and the number of those that participated. Note that the numbers of schools are somewhat smaller than in Tables 3-3 and 3-5. This is because Table 3-6 includes only schools that were to conduct reading sessions, whereas the earlier tables include all sampled fourth-grade schools.

3.6 STUDENT SAMPLE SELECTION

Schools initially sent a complete list of students to a central location in November 1991. Schools were not asked to list students in any particular order, but were asked to implement checks to ensure that all fourth-grade students were listed. Based on the total number of students on this list, called the Student Listing Form, sample line numbers were generated for student sample selection. To generate these line numbers, the sampler entered the number of students on the form and the number of mathematics and reading sessions into a calculator that had been programmed with the sampling algorithm. The calculator generated a random start that was used to systematically select the student line numbers (30 per session). To compensate for new enrollees not on the Student Listing Form, extra line numbers were generated for a supplemental sample of new students. All students were selected in those schools with grade

Table 3-5
Substitute School Counts for Grade 4

State	Double Session Substitutes	Regular Substitutes	Total
Alabama	2	27	29
Arkansas	0	13	13
California	0	16	16
Idaho	0	24	24
Indiana	0	28	28
Kentucky	0	3	3
Maine	3	53	56
Maryland	0	1	1
Massachusetts	0	15	15
Michigan	0	20	20
Minnesota	1	15	16
Mississippi	0	2	2
Missouri	0	9	9
Nebraska	0	59	59
New Hampshire	0	42	42
New Jersey	0	53	53
New Mexico	2	32	34
New York	0	28	28
North Carolina	0	5	5
North Dakota	1	46	47
Ohio	0	27	27
Oklahoma	0	15	15
Pennsylvania	0	17	17
Rhode Island	14	2	16
South Carolina	0	2	2
Tennessee	0	8	8
Texas	0	6	6
TOTAL	23	568	591

Table 3-6
Distribution of the Grade 4 Reading School Sample by State

State	Weighted Percent School Participation		Number of Schools in the Original Sample			Number of Substitute Schools for Nonparticipating Originals		Total Number of Schools That Participated
	Before Substitution	After Substitution	Total	Not Eligible	Participated	Provided	Participated	
Alabama	76.47	97.01	112	3	82	25	23	105
Arizona	99.10	99.10	107	1	106	0	0	106
Arkansas	87.04	96.45	120	2	105	12	11	116
California	91.86	97.34	115	3	103	6	6	109
Colorado	100.00	100.00	124	2	122	0	0	122
Connecticut	99.01	99.01	113	4	108	0	0	108
Delaware	92.15	92.15	56	6	44	0	0	44
Dist. of Columbia	99.33	99.33	118	4	113	0	0	113
Florida	100.00	100.00	111	1	110	0	0	110
Georgia	100.00	100.00	109	2	107	0	0	107
Guam	100.00	100.00	21	0	21	0	0	21
Hawaii	100.00	100.00	106	0	106	0	0	106
Idaho	82.25	95.52	123	1	100	19	15	115
Indiana	77.27	91.51	116	2	88	24	16	104
Iowa	100.00	100.00	133	4	129	0	0	129
Kentucky	93.77	96.58	124	3	116	3	3	119
Louisiana	100.00	100.00	115	4	111	0	0	111
Maine	57.62	71.06	141	1	76	41	20	96
Maryland	99.20	99.20	112	1	110	1	0	110
Massachusetts	86.59	96.66	123	4	103	12	11	114
Michigan	83.08	89.68	116	3	92	17	8	100
Minnesota	81.24	93.62	116	5	91	15	13	104
Mississippi	98.06	100.00	110	3	105	2	2	107
Missouri	89.56	97.14	123	6	105	9	9	114
Nebraska	76.09	86.90	161	7	106	41	15	121
New Hampshire	68.19	80.53	128	4	83	34	17	100
New Jersey	76.37	82.21	121	4	89	23	7	96
New Mexico	76.36	90.64	114	1	84	26	18	102
New York	78.20	83.69	110	0	86	21	7	93
North Carolina	95.13	99.09	118	2	111	5	5	116
North Dakota	70.36	91.48	133	3	97	33	23	120
Ohio	77.73	90.53	121	1	93	21	15	108
Oklahoma	86.34	98.09	130	0	115	14	13	128
Pennsylvania	84.61	95.50	119	0	102	17	12	114
Rhode Island	83.26	96.15	114	5	89	15	15	104
South Carolina	98.08	99.04	112	1	109	1	1	110
Tennessee	92.63	93.67	120	1	110	8	1	111
Texas	92.22	97.08	111	3	98	5	5	103
Utah	99.05	99.05	110	1	108	0	0	108
Virginia	99.00	99.00	118	4	113	0	0	113
Virgin Islands	100.00	100.00	23	0	23	0	0	23
West Virginia	100.00	100.00	144	7	137	0	0	137
Wisconsin	99.06	99.06	127	5	122	0	0	122
Wyoming	96.68	96.68	158	6	148	0	0	148

enrollment size of up to 10 percent more than the required sample size of students. This sample design was intended to give each student within the state approximately the same chance of selection.

The states where all schools were selected with certainty (Cluster Type 1 states) were treated differently. For the fourth-grade sample in Delaware and the District of Columbia, 120 students were selected, where possible. If the enrollment was lower than 120, all of the students were taken. In the territories, all of the fourth-grade students were included in the sample.

After the student sample was selected, the administrator at each school identified students who were incapable of taking the assessment either because they had an Individualized Education Plan or because they were Limited English Proficient. More details on the procedures for student exclusion are presented in the report on field procedures for the Trial State Assessment Program.

When the assessment was conducted in a given school, a count was made of the number of nonexcluded students who did not attend the session. If this number exceeded three students, the school was instructed to conduct a make-up session, to which were invited all students who were absent from the initial session.

Table 3-7 provides the distribution of the fourth-grade reading student samples and response rates by state.

Table 3-7
Distribution of the Grade 4 Reading Student Sample and Response Rates by State

State	Weighted Student Response Rate (Percent)	Number of Students			
		In Original Sample	Excluded from Sample	To Be Assessed	Actually Assessed
Alabama	95.632	2,885	153	2,684	2,571
Arizona	95.394	3,095	218	2,807	2,677
Arkansas	95.923	2,909	153	2,699	2,589
California	94.480	3,041	440	2,506	2,365
Colorado	95.211	3,275	204	3,040	2,897
Connecticut	94.699	2,914	205	2,655	2,514
Delaware	94.976	2,330	138	2,156	2,048
District of Columbia	94.323	3,033	284	2,648	2,496
Florida	94.694	3,258	296	2,925	2,767
Georgia	95.761	3,078	159	2,832	2,712
Guam	94.197	2,268	154	2,154	2,029
Hawaii	94.621	2,995	171	2,791	2,642
Idaho	95.872	2,934	112	2,789	2,674
Indiana	95.690	2,798	114	2,650	2,535
Iowa	96.371	3,006	115	2,860	2,756
Kentucky	96.145	3,007	112	2,863	2,752
Louisiana	95.690	3,159	135	2,977	2,848
Maine	94.662	2,183	123	2,038	1,916
Maryland	95.372	3,193	199	2,918	2,786
Massachusetts	95.611	2,935	224	2,663	2,545
Michigan	94.103	2,777	136	2,615	2,437
Minnesota	95.655	2,895	117	2,741	2,589
Mississippi	96.580	2,981	150	2,753	2,657
Missouri	95.457	2,834	124	2,686	2,562
North Carolina	96.371	3,128	136	2,991	2,883
North Dakota	97.116	2,275	48	2,222	2,158
Nebraska	95.810	2,648	126	2,496	2,364
New Hampshire	96.093	2,554	115	2,417	2,239
New Jersey	95.606	2,510	139	2,342	2,239
New Mexico	94.628	2,852	214	2,508	2,305
New York	94.558	2,594	149	2,418	2,285
Ohio	95.560	2,910	179	2,704	2,580
Oklahoma	84.945	2,936	240	2,658	2,251
Pennsylvania	95.485	3,071	122	2,941	2,805
Rhode Island	95.274	2,764	192	2,464	2,347
South Carolina	96.435	3,083	170	2,857	2,758
Tennessee	95.115	3,047	141	2,874	2,734
Texas	96.053	2,987	252	2,678	2,571
Utah	96.399	3,139	140	2,934	2,829
Virginia	95.620	3,128	199	2,914	2,786
Virgin Islands	96.815	932	33	911	882
West Virginia	95.961	3,009	152	2,848	2,733
Wisconsin	95.984	3,049	199	2,827	2,712
Wyoming	95.848	3,046	124	2,894	2,775

Chapter 4

STATE AND SCHOOL COOPERATION AND FIELD ADMINISTRATION

Nancy Caldwell

Westat, Inc.

4.1 OVERVIEW

By volunteering to participate in the Trial State Assessment and in the field test that preceded it, each state assumed responsibility for securing the cooperation of the schools sampled by NAEP. The participating states were responsible for the actual administration of the 1992 Trial State Assessment at the school level. For the field test in 1991, however, individual states could choose to have NAEP administer the entire program. This chapter describes state and school cooperation and field administration procedures for both the field test and the 1992 program. Section 4.2 presents information on the field test in 1991, while section 4.3 focuses on the 1992 Trial State Assessment.

4.2 THE FIELD TEST

4.2.1 Conduct of the Field Test

In preparation for the 1992 state and national assessment programs, a field test of the forms, procedures, and booklet items was held in early 1991. The field test also gave states an opportunity to learn about their responsibilities for the new aspects of the Trial State Assessment.

In June 1990, letters were sent from the U.S. Department of Education to all Chief State School Officers inviting them to participate in the field test of materials and procedures for 1992. Since the fourth grade had not been assessed as part of the Trial State Assessment before, states were given the option of conducting the field test themselves for this grade. At the eighth grade, only states that had not participated in the 1990 assessment were given the option of conducting the field test themselves. In an effort to secure the participation of more schools and to lessen the burden of participation on the states, ETS and Westat offered to perform all of the work involved, including communicating with school staff, sampling, and administering the assessment.

Twenty-four jurisdictions decided to participate in the field test. Each participating jurisdiction was asked to appoint a state coordinator to secure the cooperation of sampled schools, and to be the liaison between NAEP/Westat staff and the participating schools.

As described in Chapter 3, the state coordinator for each state was sent the names of approximately 30 pairs of selected schools and requested to secure the cooperation of one school from each pair. This process had been used successfully in the field test for 1990, and was again successful in the field test for 1992. In total, 664 schools agreed to participate in the field test; in 662 of these schools assessment sessions were conducted.

Twenty-one of the jurisdictions decided to have NAEP administer all field test sessions. In these jurisdictions, the state coordinator secured the cooperation of the selected schools and then Westat contacted the schools, confirmed the schedule and arrangements, selected the student samples, and conducted the assessment sessions.

Three states—Florida, Kentucky, and Wisconsin—chose to have school staff (assessment administrators) conduct the fourth-grade assessments, while none of the jurisdictions elected to conduct the eighth-grade assessments themselves. Although the three states were responsible for the actual administration at the school level, Westat was responsible for developing the administration materials and procedures and for training state staff. Two training sessions were conducted by Westat home office staff in each of the three states during mid-January. All assessment administrators received a manual before attending one of these training sessions. The training program consisted of a video presentation, scripted lecture and training exercises.

In January 1991, Westat field supervisors selected the student sample for each school and prepared an Administration Schedule (roster) of the sampled students. The Administration Schedule was sent by the state coordinator to the school two weeks before the scheduled assessment date. The other assessment materials were shipped by NCS to arrive two weeks before the scheduled assessment date. Upon receiving the Administration Schedule and the assessment materials, the assessment administrator followed NAEP procedures to select an additional sample of newly enrolled students, identify students who were not capable of participating in the assessment, and prepare assessment questionnaires.

On assessment day, the field supervisor observed the assessment and queried the assessment administrator about the session, procedures, and materials. Supervisors used an Observation Form to record information about the major events related to the assessment and the assessment administrators' opinions and comments.

4.2.2 Results of the Field Test

The overall desired student participation level for the field test was determined from the goal of obtaining 300 student responses for each item to be used in the national assessment and 500 student responses for each item to be used in the Trial State Assessment. Depending on the size of the school, the school's sample numbered approximately 30 to 60 students, who were assessed in either one or two sessions.

Given these goals, the overall desired student participation in both the national and Trial State components of the field test was 22,600 students. In actuality, 24,910 students, or about 10 percent more than required, were assessed.

The field testing of materials and procedures at the fourth-grade level for the Trial State Assessment in the three states provided useful information for NAEP staff in preparation for 1992. While the sessions went well and 80 to 90 percent of assessment administrators thought that the training session, the manuals, and the assessment materials worked well, the administrators did make many suggestions for improving these materials and procedures for the 1992 assessment program.

4.3 THE 1992 TRIAL STATE ASSESSMENT

Forty-one states, the District of Columbia, and two territories volunteered for the 1992 Trial State Assessment. This is a net increase of four jurisdictions over 1990, with seven newly participating in 1992 and three that were in the 1990 assessment deciding not to participate in 1992. Figure 4-1 identifies the jurisdictions participating in each of the two assessment years. (Similar information is presented in table form in Chapter 1.) As with the field test, each jurisdiction designated a state coordinator to oversee all assessment activities in the state.

Two states—Illinois and Washington—had agreed to participate in the 1992 Trial State Assessment, but dropped out before the assessment began, primarily due to a lack of success in getting schools in their states to participate. This followed a letter from NCES recommending that states obtain at least a 70 percent school cooperation rate in order to meet the guidelines for participation.

4.3.1 Overview of Responsibilities

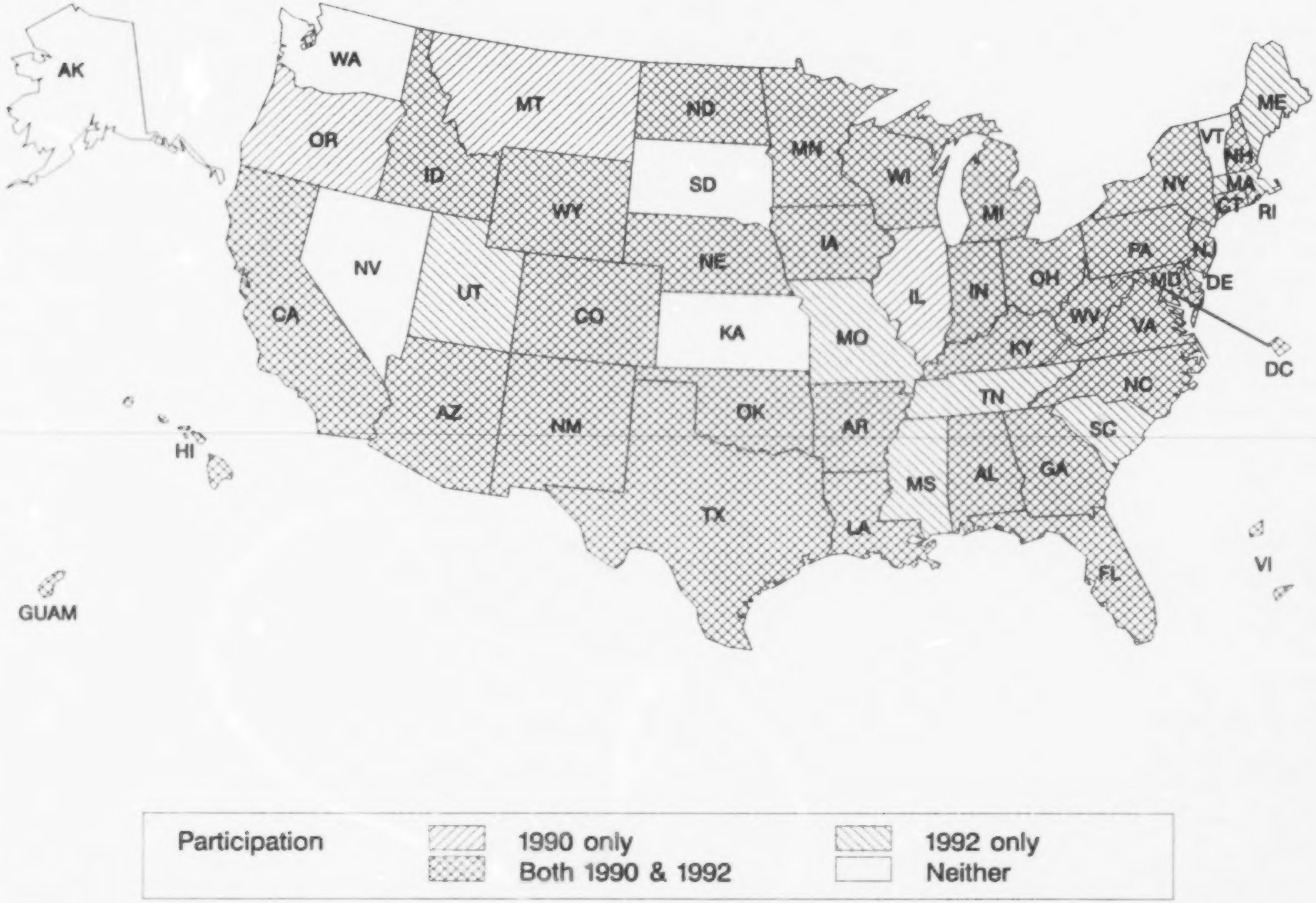
The data collection for the 1992 Trial State Assessment involved a collaborative effort between the participating states and the NAEP contractors, especially Westat, the field administration contractor. Westat's responsibilities included

- selecting the sample of schools and students for each participating state;
- developing the administration procedures and manuals;
- training the state personnel who would conduct the assessments; and
- conducting an extensive quality assurance program.

Each jurisdiction volunteering to participate in the 1992 program was asked to appoint a state coordinator. In general, the state coordinator was the liaison between NAEP/Westat staff and the participating schools. In particular, the state coordinator was asked to

- gain the cooperation of the selected schools;
- assist in the development of the assessment schedule;
- receive the lists of all grade-eligible students from the schools;

Figure 4-1
Participating Jurisdictions, 1990 and 1992 Trial State Assessments


AK
WA
OR
MT
ND
MN
ME
VT
NH
MA
CT
RI
ID
SD
WI
MI
NY
WY
IA
PA
NJ
NV
NE
OH
MD
DE
UT
CA
CO
IL
IN
WV
KA
MO
KY
VA
DC
TN
NC
AZ
NM
OK
AR
SC
HI
MS
AL
GA
TX
LA
FL
VI
GUAM
Participation
1990 only
Both 1990 & 1992
1992 only
Neither

- coordinate the flow of information between the schools and the NAEP contractors;
- provide space for the state supervisor to use when sampling;
- notify assessment administrators about training and send them their manuals; and
- send the lists of sampled students to the schools.

At the local school level, an assessment administrator was responsible for preparing for and conducting the assessment session(s) in one or more schools. These individuals were usually school or district staff and were trained by Westat staff. The assessment administrator's responsibilities included

- receiving the list of sampled students from the state coordinator;
- identifying sampled students who should be excluded;
- distributing assessment questionnaires to appropriate school staff;
- notifying sampled students and their teachers;
- administering the assessment session;
- completing assessment forms; and
- preparing the assessment materials for shipment.

Westat hired and trained six field managers and 44 state supervisors, one for each jurisdiction. Each field manager was responsible for working with the state coordinators of seven to eight states and for overseeing assessment activities. The primary tasks of the field managers were to

- obtain information about cooperation and scheduling;
- make sure the arrangements for the assessments were set and assessment administrators identified; and
- schedule the assessment administrators training sessions.

The primary tasks of the state supervisors were to

- select the sample of students to be assessed;
- conduct in-person assessment administrator training sessions; and
- coordinate the monitoring of the assessment sessions and makeup sessions.

Westat also hired and trained an average of eight quality control monitors in each state to monitor 50 percent of the assessment sessions.

4.3.2 Schedule of Data Collection Activities

May 15, 1991	Westat sent the samples of schools selected for the National and Trial State Assessment to the state coordinators.
Early August, 1991	Westat field managers visited each state to explain the computerized State Coordinator System, which could be used to keep track of assessment-related activities. Westat distributed Student Listing Forms, Principal Questionnaires, and the list of the schools selected for the Trial State Assessment updated with a suggested week of assessment and number and type of sessions.
May-November, 1991	State coordinators obtained cooperation from districts and schools. State coordinators reported participation status to Westat field managers via printed lists or computer files. State coordinators sent Student Listing Forms, Supplemental Student Listing Forms, and Principal Questionnaires to participating schools.
October-November, 1991	Westat selected substitutes for refusals and sent them to state coordinators. States reporting the participation status of all schools by October 15 received substitutes for refusals by October 31. States reporting by October 31 received substitutes by November 15.
November 14-17, 1991	State supervisor training sessions were held.
December 2-20, 1991	NAEP state supervisors visited state coordinators to select student samples and prepare Administration Schedules listing the students selected for each session. Westat provided the schedule of training sessions and copies of the Manual for assessment administrators to state coordinators for distribution.
December 2, 1991-January 10, 1992	State coordinators notified assessment administrators of the date and time of training and sent each a copy of the *Manual for Assessment Administrators*.
January 3-10, 1992	Quality control monitor training sessions were held.
January 9-31, 1992	Assessment administrator training sessions were held.

January 20- February 14, 1992	State coordinators sent Administration Schedules to each school two weeks before the scheduled assessment date.
February 3-28, 1992	Assessments were conducted. Unannounced visits were made by quality control monitors to a predetermined 50 percent of the sessions.
March 2-6, 1992	Makeup sessions were held as necessary.

4.3.3 Preparations for the Trial State Assessment

The focal point of the schedule for the Trial State Assessment was the period between February 3-28, 1992, when the assessments were conducted in the schools. However, as with any undertaking of this magnitude, the project required many months of planning and preparation.

Westat selected the samples of schools according to the procedures described in Chapter 3. On May 15, 1991, lists of these selected schools and other materials describing the Trial State Assessment Program were sent to state coordinators. This mailing took place about two months earlier than in the 1990 assessment because state coordinators had requested more time to contact districts and schools and schedule the assessments. Most state coordinators also preferred that NAEP provide a suggested assessment date for each school. School listings were updated with this information and were sent to the state coordinators, along with other descriptive materials and forms, in early August.

State coordinators also were given the option of receiving the school information in the form of a computer database with accompanying management information software. This system enabled the state coordinators to keep track of the cooperating schools, the assessment schedule, the training schedule, and the assessment administrators. Coordinators could choose to receive a laptop computer and printer or to have the system installed on their own computer. Westat field managers traveled to the state offices to explain the computer system to the state coordinators and their staff. All but one state coordinator chose to receive the system.

Six of the most experienced NAEP supervisors were chosen to be field managers, the primary link between NAEP and the state coordinators. In mid-August, the field managers visited offices of the state coordinators to explain the computerized system to state staff. The field managers kept in frequent contact with the state coordinators as the state coordinators secured the cooperation of the selected schools and established the assessment schedule.

The field managers used the same computer system as the state coordinators to keep track of the schools and schedule. The state coordinators sent updates either via computer disks, by telephone, or in print to their field manager, who then entered in the information into the system. Weekly transmissions were made from the field manager to Westat.

The state coordinators' first task was to secure the participation of the selected schools. States that had determined the cooperation status of all selected schools by October 15 were sent a list of potential replacements for refusals by October 31. States that reported by October

31 received a list of potential substitutes by November 15. Both printed lists and computer files of substitute schools were transmitted to the field managers and state coordinators. (See Chapter 3 for more details about school substitution.)

In mid-November, Westat hired one state supervisor for each participating state. The state supervisors attended a training session held in the Washington, DC, area between November 14-17, 1991. This training session focused on the state supervisors' immediate tasks—selecting the student samples and hiring quality control monitors. State supervisors also were given the training script and materials for the assessment administrators' training sessions they would conduct in January so they could begin to become familiar with these materials.

The state supervisors' first task after training was to complete the selection of the sample of students who were to be assessed in each school. All participating schools were asked to send a list of their grade-eligible students to the state coordinator by November 15. Sample selection activities were conducted in the state coordinator's office unless the state coordinator preferred that the lists be taken to another location.

Using a preprogrammed calculator, the supervisors generally selected a sample of 30 students per session type per school. The exceptions to this were small schools and states with fewer than the necessary 125 fourth-grade schools. In the states with fewer schools, larger student samples were required from schools that participated.

After the sample was selected, the supervisor completed an Administration Schedule for each session, listing the students to be assessed. The Administration Schedules for each school were put into an envelope and given to the state coordinator to send to the school two weeks before the schedule assessment date. Included in the envelope were instructions for sampling students who had enrolled at the schools since the creation of the original list used in sampling.

During the period from mid-November through December, the state supervisors also recruited and hired quality control monitors to work in their states. It was the quality control monitor's job to observe the sessions designated to be monitored, complete an observation form on each session and to intervene when the correct procedures were not followed. In each state, half of the sessions were designated to be monitored. This information was known only to contractor staff; it was not on any of the listings provided to state staff.

Approximately 400 quality control monitors were trained in two training sessions held during January 3-6 and 7-10, 1992. The first day of each training session was devoted to a presentation of the assessment administrators training program by the state supervisors, which not only gave the quality control monitors an understanding of what assessment administrators were expected to do, but gave state supervisors an opportunity to practice presenting the training program. The remaining days of the training sessions were spent reviewing the quality control monitor observation form and the role and responsibilities of the quality control monitors.

Almost immediately after the quality control monitor training sessions, the supervisors began conducting the assessment administrator training sessions. Each quality control monitor attended several of these training sessions, to assist the state supervisor and to become thoroughly familiar with the assessment administrator's responsibilities. Almost 10,000 persons who were to be assessment administrators were trained in about 500 training sessions across the nation.

To ensure uniformity in the training sessions, Westat developed a highly structured program involving a script for trainers, a videotape, and a training example to be completed by the trainees. The supervisors were instructed to read the script verbatim as they proceeded through the training, ensuring that each trainee received the same information. The script was supplemented by the use of overhead transparencies, displaying the various forms that were to be used and enabling the trainer to demonstrate how they were to be filled out.

The videotape, similar to the one used in the 1990 Trial State Assessment, was developed by Westat to provide background for the study and to simulate the various steps of the assessment that would be repeated by the assessment administrators. The portions of the videotape depicting the actual assessment had been taped in a classroom with students in attendance to closely simulate an actual assessment session. The videotape was divided into sections with breaks for review by the trainer and practice for the trainees.

The final component of the presentation was the "Training Example." This consisted of a set of exercises keyed to each part of the training package. A portion of the videotape was shown and then reviewed by the trainer following the script. Then, exercises related to that material were completed by the trainees before the next subject was discussed.

The entire training session generally ran for about three and one-half hours. Sessions usually began in the morning and ended with lunch. In 1990, the training sessions had generally lasted about five to six hours. Responding to requests from state coordinators and assessment administrators, Westat trimmed the training session to half of a day.

All of the information presented in the training session was included in the *Manual for Assessment Administrators*, developed by Westat. Copies of the manuals were sent by Westat to the state coordinators at the beginning of December so that they could be distributed to the assessment administrators before the training sessions.

4.3.4 Monitoring of Assessment Activities

Two weeks prior to the scheduled assessment date, the assessment administrator received the Administration Schedule and assessment questionnaires and materials. Five days before the assessment, the quality control monitor made a call to the assessment administrator and recorded the results of the call on the Observation Form. Most of the questions asked in the pre-assessment call were designed to gauge whether the assessment administrator had received all materials needed and was prepared for the session.

Pre-assessment calls were made to all schools regardless of whether they were to be monitored. If the sessions in a school were not observed, the quality control monitor called the assessment administrator three days after the assessment to find out how the session went, to obtain the assessment administrator's impressions of the manual, training, and materials and to ensure that all post-assessment activities had been completed.

If the sessions in a school were to be monitored, the quality control monitor was to arrive at the school one hour before the scheduled beginning of the assessment to observe preparations for the assessment. To ensure the confidentiality of the assessment items, the booklets were packaged in shrink-wrapped bundles and were not to be opened until the quality control monitor arrived or 45 minutes before the session began, whichever occurred first.

In addition to observing the opening of the bundles, the quality control monitor used the Observation Form to check that the following had been done correctly: sampling newly enrolled students, reading the script, distributing and collecting assessment materials, timing the booklet sections, answering questions from students, and preparing assessment materials for shipment.

After the assessment was over, the quality control monitor obtained the assessment administrator's opinions of how the session went and how well the materials and forms worked.

If four or more students were absent from the session, a makeup session was to be held. If the original session had been monitored, the makeup session was also monitored. This required coordination of scheduling between the quality control monitor and assessment administrator.

4.3.5 School and Student Participation

Table 4-1 shows the results of the state coordinators' efforts to gain the cooperation of the selected schools. Overall, 4,921 fourth-grade schools participated in the 1992 Trial State Assessment. This is about 88 percent (unweighted) of the eligible schools in the original sample at each grade and about 95 percent (unweighted) of the sample after substitution.

Table 4-1
Fourth-grade School Participation, 1992 Trial State Assessment

Status	Grade 4
Schools in original sample	5356
Schools not eligible (e.g., closed, no grade 4/8)	152
Eligible schools in original sample	5204
Noncooperating (e.g., school, district, state refusal)	605
Participating	4599
Substitutes provided for noncooperating schools	501
Participating substitutes	322
Total schools participating after substitution	4921

Participation results for students in the 1992 Trial State Assessment in reading are given in Table 4-2. Approximately 129,000 students were sampled. As can be seen from the table, the original sample, which was selected by the NAEP state supervisors, comprised about 125,000 of this number. The original sample size was increased somewhat after the supplemental samples had been drawn (from students newly enrolled since the creation of the original lists).

Table 4-2
Student Participation in the 1992 Trial State Assessment of Reading

Status	Grade 4 Reading
Sampled	129,322
Original sample	125,445
Supplemental sample	3,877
Withdrawn	5,668
Excluded	7,306
To be assessed	116,348
Assessed	110,852
Initial sessions	110,469
Makeup sessions	383

Assessment administrators removed some students from the total sample according to NAEP criteria: first, those students who had left their schools since the time that they were sampled (withdrawn); then, those judged incapable of participating meaningfully in the assessment by school staff (excluded). A student could be excluded if she or he either had an Individualized Education Plan (IEP) or was classified as Limited English Proficient (LEP), was incapable of participating meaningfully, and met certain other criteria.

These exclusions left 116,348 fourth graders to be assessed in reading. Of these, 110,852 were actually assessed, yielding an unweighted student participation rate of 95.3 percent.

4.3.6 Results of the Observations

During the assessment sessions, the quality control monitors were to note instances when the assessment administrators deviated from the prescribed procedures and whether any of these deviations were serious enough to warrant their intervention. Quality control monitors reported no instances where there were serious breaches of the procedures or major problems that would question the validity of the assessment.

Deviation from prescribed procedures occurred most often in the administrator's reading of the script that introduced the assessment and provided the directions. Even so, in at least 90

BEST COPY AVAILABLE

percent of the observed sessions the assessment administrator read the script verbatim or with only slight deviations. Examples of major deviations included skipping sections of the script, adding substantially to the script, and forgetting to pass out materials at the appropriate times. The quality control monitor intervened in these instances.

Most of the other procedures that could have had some bearing on the validity of the results were adhered to very well by the assessment administrators. In 99 percent of the observed sessions, the assessment administrators opened the bundles of booklets at the appropriate time and handled questions from the students correctly. Ninety-nine percent of the fourth-grade sessions were timed correctly.

After the assessment session was over, assessment administrators were asked how they thought the assessment went and whether they had any comments or suggestions. Overall, assessment administrators stated that they thought 98 to 99 percent of the sessions went very well or satisfactorily.

Assessment administrators reported that 79 percent of the fourth-grade reading sessions went very well, with a higher percentage of monitored sessions (81%) than unmonitored sessions (77%) reported as going very well.

Comments about the assessment materials and procedures were generally favorable. Criticisms or suggestions included that there were too many forms and too much paperwork; coding the booklet covers was tedious and problematic for students; and schools needed more information about NAEP and assessment results.

In addition to these interviews, Westat sent a debriefing form to all of the NAEP state supervisors and met in person with half of them. This meeting produced suggestions for future assessments, especially many minor changes in the procedures, materials and training plans. In addition, the state supervisors recommended that district and particularly school staff receive more information describing the background and objectives of NAEP and the Trial State Assessments. They also stated that many school staff were very interested in results for their students, or at least summary results for their state.

State coordinators were also sent a questionnaire about their experiences, suggestions, and comments. State coordinators from 39 of the participating states and territories responded. All of the 35 state coordinators responding to the question "How did the assessments go in your state?" said "Very well" to "Fairly well." They also commented favorably on the training package and other materials. Like the assessment administrators, the state coordinators criticized the amount of work required to prepare for the assessments. They made many other suggestions about the computerized data system, sampling procedures, training program, and design of the assessment. All of these suggestions will be reviewed as future assessments are planned.

The results of the assessment and comments from assessment administrators and state coordinators were summarized in a report presented to the NAEP Network on May 11, 1992. In mid-August, each participating state and territory received a summary of its participation data, data collection activities, results of the assessment, and assessment administrators' comments.

Chapter 5

PROCESSING AND SCORING ASSESSMENT MATERIALS

Dianne Smrdel, Linda Reynolds, and Brad Thayer

National Computer Systems

5.1 OVERVIEW

This chapter describes the printing, distribution, receipt, processing, scoring, and final disposition of materials for the reading portion of the Trial State Assessment. The scope of the effort required by National Computer Systems (NCS) to process the materials is evidenced by the following:

- Prior to the assessment, 15,528 bundles of assessment booklets were created and distributed to approximately 9,000 schools.
- One booklet was processed for each of the approximately 111,000 students assessed; 35,800 questionnaires were received and processed; and about 1.7 million student responses from 43 constructed-response items were professionally scored.
- In all, approximately 3.6 million double-sided pages from test booklets and questionnaires were optically scanned.

Throughout the processing, the NCS Process Control System and Workflow Management System were used to track, audit, edit, and resolve characters of information. A quality control sample of characters of transcribed data was selected and compared to the actual responses in the assessment booklets.

The volume of collected data and the complexity of the Trial State Assessment processing design, with its spiraled distribution of booklets, as well as the concurrent administration of this assessment and the national assessments, required the enhancement and implementation of flexible, innovatively designed processing programs and a sophisticated Process Control System. This system, developed for the 1990 assessments, allowed an integration of data entry and workflow management systems, including carefully planned and delineated editing, quality control, and auditing procedures.

The magnitude of the effort is apparent when considering that the activities described in this chapter were completed concurrently with the processing of the national assessments, that all processing activities were completed within 10 weeks, and that an estimated accuracy rate of fewer than five errors for every 10,000 characters of information was achieved.

Several major changes in materials processing were made from 1990, including the conversion of all documents to scannable form, the tailoring of shipments to the individual size and requirements of schools, and the reorganization of the process flow to conduct constructed-response scoring after all machine scoring and data verification processes were complete, allowing NCS to provide Westat and ETS with demographic and cognitive data at an earlier date.

5.2 PROCESS CONTROL SYSTEM

NCS maintains a Process Control System consisting of numerous specialized programs and processes to accommodate the unique demands of concurrent assessment processing and a unified ETS/NCS system integration. The Process Control System, which was developed for the 1990 assessment, was necessary to maintaining control of all shipments of materials to the field, of all receipt from the field, and of any work in progress. The system is a unique combination of several reporting systems currently in use at NCS, along with some application-specific processes. These systems are the Workflow Management System, the Bundle Assembly Quality Control System, the Outbound Mail Management System, and the On-line Inventory Control system. Data were collected from these systems and recorded in the file called the "NAEP Process Control System." Additional information was directly entered into the Process Control System.

5.3 WORKFLOW MANAGEMENT SYSTEM

The functions of the Workflow Management System are to keep track of where the production work is and where it should be and to collect data for status reporting, forecasting, and other ancillary subsystems. The primary purpose of the Workflow Management System is used to analyze the current workload by project across all work stations.

The data processing and control systems are determined to a large extent by the type of documents processed. For the Trial State Assessment, only machine-scannable assessment booklets and answer documents were used to collect student responses. The three questionnaires that were used to collect data about school characteristics, teachers associated with sampled students, and students excluded from the assessment were also scannable documents.

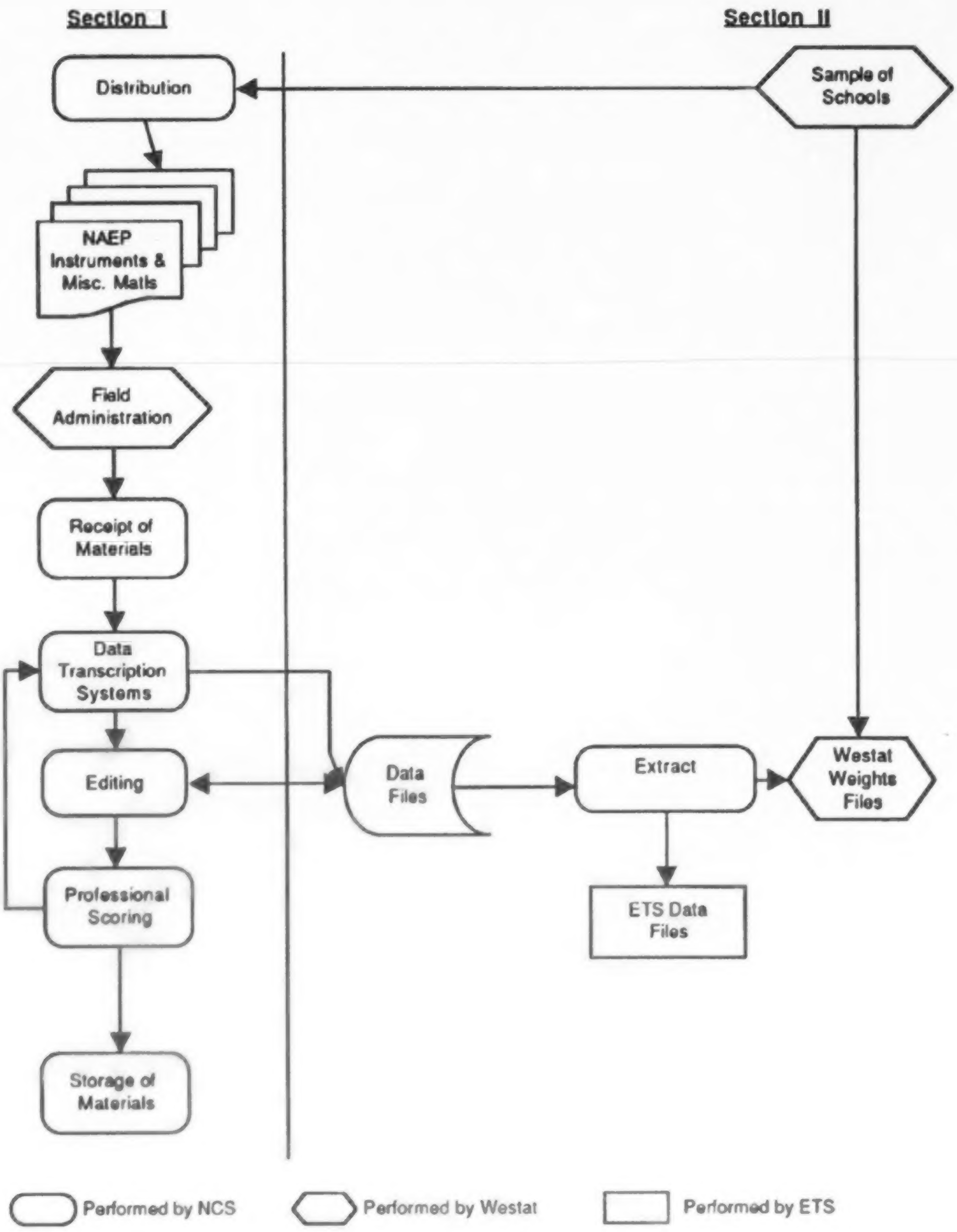
5.4 PROCESS FLOW OF NAEP MATERIALS AND DATABASE CREATION

Figure 5-1 shows the conceptual framework of processes that were used both for the Trial State Assessment materials and for the national NAEP materials.

Section I of Figure 5-1 depicts the flow of NAEP's printed materials. Information from the Administration Schedule and Packing List was used to control the processing of materials.

Figure 5-1

Data Flow Overview, 1992 Trial State Assessment


Section I
Section II
Distribution
Sample of Schools
NAEP Instruments & Misc. Matls
Field Administration
Receipt of Materials
Data Transcription Systems
Editing
Professional Scoring
Storage of Materials
Data Files
Extract
Westat Weights Files
ETS Data Files
Performed by NCS
Performed by Westat
Performed by ETS

The figure follows the path of each assessment instrument—Student Test Booklets, School Characteristics and Policies Questionnaires, Teacher Questionnaires, Excluded Student Questionnaires, Packing List, and Administration Schedules—as they were tracked through the appropriate processes that resulted in the final integrated NAEP database.

The remainder of this chapter provides an overview of the materials processing activities as shown in Section I of Figure 5-1 and detailed in Figure 5-2. Section II of Figure 5-1 depicts the evolution of the NAEP/NCS database from the transcribed data to the final files, provided to Westat for creation of weights and to ETS for analysis and reporting.

The 1992 NAEP data collection resulted in six classes of data files (student, school, teacher, excluded student, sampling weight, and item information files). The structure and internal data format of the 1992 NAEP database was a continuation of the integrated design originally developed by ETS in 1983.

5.5 MATERIALS DISTRIBUTION

The use of bar code technology in document control was introduced to NAEP by NCS in the 1990 assessment; its use continued in 1992. Bar codes were applied to the front cover of the documents. The bar code consisted of the two-digit booklet number, a five-digit sequential number, and a check digit.

The booklets were spiraled into 16 unique bundles consisting of 11 booklets in a set pattern. A header sheet was attached to each bundle that indicated the assessment type, bundle type, bundle number, and a list of the booklet types to be included in the bundle.

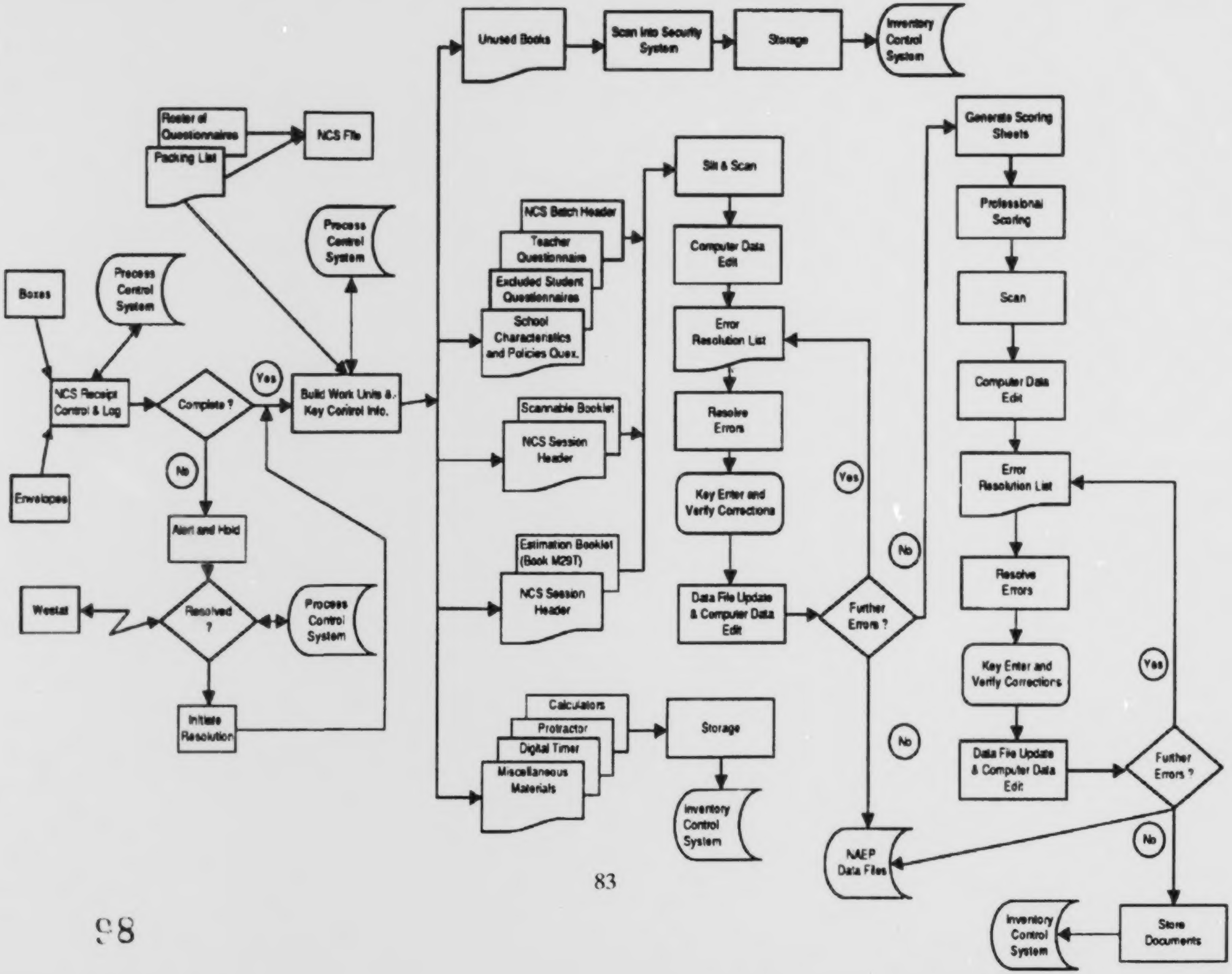
The bundle numbers on the header sheet were created to identify the type of bundle. All bundles were then passed under a scanner programmed to interpret this type of bar code and the file of scanned barcodes was transferred from the scanner to the mainframe. A computer program compared the bundle type expected to the one actually scanned after the header and verified that there were 11 booklets in each bundle. Any discrepancies were printed on an error listing forwarded to the Packaging Department, where the error was corrected and the bundle was again read into the system for another quality control check. This process was repeated until all bundles were correct.

The bundles were shrink-wrapped in clear plastic. A bright label was placed over the cross of the straps that read "Do Not Open Until 45 Minutes Before Testing." Following this, bundles were ready for assignment and distribution.

When packing lists for distribution of materials were created from the Materials Distribution System, a second and more detailed bundle slip was produced. This bundle slip indicated the same information as the slip wrapped with the bundle, in addition to the school number and the complete booklet ID numbers of the booklets within that bundle. This allowed the assessment administrators to pre-assign booklets for their sessions.

Figure 5-2

Materials Processing Flow, 1992 Trial State Assessment


Unused Books
Scan into Security System
Storage
Inventory Control System
Roster of Questionnaires
Packing List
NCS File
Process Control System
Boxes
Process Control System
NCS Receipt Control & Log
Envelopes
Complete ?
Yes
No
Build Work Units & Key Control Info.
Alert and Hold
Westat
Resolved ?
Process Control System
Initiate Resolution
NCS Batch Header
Teacher Questionnaire
Excluded Student Questionnaires
School Characteristics and Policies Quex.
Scannable Booklet
NCS Session Header
Estimation Booklet (Book M29T)
NCS Session Header
Calculators
Protractor
Digital Timer
Miscellaneous Materials
Storage
Inventory Control System
Slit & Scan
Computer Data Edit
Error Resolution List
Resolve Errors
Key Enter and Verify Corrections
Data File Update & Computer Data Edit
Further Errors ?
Yes
No
No
NAEP Data Files
Generate Scoring Sheets
Professional Scoring
Scan
Computer Data Edit
Error Resolution List
Resolve Errors
Key Enter and Verify Corrections
Data File Update & Computer Data Edit
Further Errors ?
Yes
No
Store Documents
Inventory Control System

The timing of the shipments of these materials to the participating schools was critical, since the shipments needed to be in the school at least one week but not more than two weeks prior to testing.

Each school conducted at least one session; some conducted more than one. The materials needed for a school to conduct all of its mathematics sessions were sent in one shipment. The booklets for the fourth-grade reading session(s) were boxed separately in the same shipment. In 1990, each session's materials had been shipped independently. Although this change in shipment practice eliminated the option to pre-assemble many materials, it did cause less confusion within the schools.

Some materials were distributed per school; others were distributed per session. Materials issued for each reading assessment session were:

Bundle(s) of 11 assessment booklets (based on sample count)
1 Pad of appointment cards
1 Return postage paid label
1 Post-it note pad
1 Shipping tape
5 Excluded Student Questionnaires
5 Teacher Questionnaires

Those materials distributed to each school were:

2 Roster of Questionnaires
2 Assessment Notifications
1 Pre-addressed envelope
1 School Characteristics and Policies Questionnaire
1 Pre-addressed box

Shipments were sent according to the week of assessment. Some schools found they needed extra quantities of materials (i.e., more excluded student questionnaires or more teacher questionnaires) and calls were received requesting these additional materials.

Aiding in the security of the shipments was the decision to send all shipments, whenever possible, through Airborne. NCS is connected to the Airborne system through computer link thus expediting tracing of any misdirected shipments. This system provides the date and time of delivery as well as the name of the person who signed for the shipment. All shipments were recorded in the Airborne Libra system. If a shipment had to be sent by UPS or the U.S. Postal Service, this information was also recorded and transferred to the mainframe.

5.6 PROCESSING ASSESSMENT MATERIALS

The materials from each session were to be returned to NCS in the same box in which they were originally mailed. It was the responsibility of the assessment administrator in the unmonitored schools and the quality control monitor in the monitored schools to repackage the items in the proper order, complete all paperwork and return the shipment through the U.S. Postal Service, using the postage-paid label provided.

With approximately 9,000 individual shipments arriving over a four-week period, it was necessary to devise a system that would quickly acknowledge receipt of a school's material. A label applied to the outside of the box by the NCS packaging department contained a bar code which indicated the school number and the project number. When the shipment arrived at NCS, the bar code was read and the shipment forwarded to the receiving area. The file was then transferred to the mainframe through a PC link and a computer program was used to apply the shipment receipt date to the appropriate school within the Process Control System. This provided current status of shipments received regardless of any processing backlog. This information was then transferred electronically to Westat. The status of the administration was checked and in some cases a trace was initiated on the shipment.

Receiving personnel also checked the shipment to verify that the contents of the box matched the school and session indicated on the label. Each shipment was checked for completeness and accuracy, regardless of whether it was monitored or unmonitored.

The materials were checked against the Packing List (see Figure 5-3) to verify that all materials were returned. If any discrepancies were found, an alert was issued. If all assessment instruments were returned, processing continued.

Each booklet and Excluded Student Questionnaire was verified against the Administration Schedule. This included verification of all counts of booklets returned and the matching of information on the front cover of the booklets to that on the Administration Schedule. If any discrepancy was discovered, an alert was issued.

After the contents of the shipment had been identified and verified, the information from the Administration Schedule was entered into the Process Control System. That information included school number, session code, counts of the number of students in original sample, supplemental sample, total sample, withdrawn, excluded, to be assessed, absent, original assessed, assessed in makeup and total assessed. If a makeup session was expected, an information alert was issued to facilitate tracking. The control counts were used by NCS for verification of processing counts. This information was also transferred electronically to Westat on a weekly basis to be used to produce participation statistics for the states.

If quantities and individual information matched, the booklets were organized into work units and batched for processing. The processing flow was changed in 1992, resulting in the completion of the machine scoring prior to the constructed-response scoring. Each batch, consisting of multiple sessions, was assigned a unique batch number. The batch number was entered on the Workflow Management System, facilitating the internal tracking of the session and allowing departmental resource planning. A scannable session header, included in the shipment from the school, was coded with the session code and placed on top of the stack of documents. All student documents were forwarded to machine scanning functions. Control documents were forwarded to appropriate record filing systems.

The excluded student questionnaires and teacher questionnaires were compared to the Roster of Questionnaires and the Administration Schedule to verify demographic information. Some questionnaires may not have been available for return with the shipment. These were returned to NCS at a later date in an envelope provided for that purpose. If the Excluded Student Questionnaire was not returned with the shipment of booklets, a record containing all demographic information on that student from the Administration Schedule was entered into the Process Control System. If the questionnaire was subsequently returned, this record was deleted. Otherwise, the record was provided to Westat for use in the weighting process.

Each school characteristics and policies questionnaire was compared with the Roster of Questionnaires and the school number was verified to match all other materials in the shipment. As with the other questionnaires, this document may not have been returned with the shipment and could also be returned in the supplemental envelope. There was no additional effort made to collect or report information on unreturned school questionnaires.

All assessed and absent students were assigned a test booklet. To indicate an absence, the "A" bubble in the Administration Code column on the front cover of the booklet was gridded. The booklet was then processed with assessed student booklets to maintain session integrity.

The Packing List (Figure 5-3) was used by the schools to account for all materials received from and returned to NCS. Any discrepancies in quantities received or returned to NCS were indicated. Also indicated was whether a makeup session was to be held, the date of scheduled makeup, the number of students involved, and the quantities of materials being held for later return.

The Administration Schedule contained the demographic characteristics of the students selected for the assessment. This information included the sex, race/ethnicity, birth date, and IEP/LEP indicators. The booklet number of the student selected was recorded on the Administration Schedule during the assessment process, and the demographic information was transferred to the booklet covers by either the student or the assessment administrator.

The demographics of the sampled students who did not participate in the assessment (exclusions and absentees) were provided to Westat to be used to adjust the sampling weights of the students who did participate. The excluded student information was obtained from the excluded student questionnaire or provided on a file for those not returned to NCS. The absent student information was taken from the front cover of the booklet that was assigned prior to the start of the assessment. This procedure eliminated the need for an additional form for absent students.

For the Rosters of Questionnaires, two numbers were entered for each type of questionnaire: number of questionnaires expected and number actually received. The Packing List, Administration Schedule, and Roster of Questionnaires were forwarded to the operations coordinator and filed by school within state for future reference.

5.7 PROFESSIONAL SCORING

The 1992 Trial State Assessment in reading for grade 4 contained short constructed-response and extended constructed-response items. These items were administered in scannable assessment booklets that were identical to the reading booklets used in the grade 4 national assessment.

Figure 5-3

Packing List, 1992 Trial State Assessment

Packing List

Seq: 00001

NAEP - 1992

Section I. Materials:	# Received from NAEP		# of Items Returned to NAEP	Section III. Held for Makeup
Math Grade 4 Booklets (459-460)	1 bundle(s)	used	________ea	________
		unused	________ea	
Reading Grade 4 booklets (923-924)	1 bundle(s)	used	________ea	________
		unused	________ea	
Cassette Tape - M29T	01		________	________
Timer	02		________	________
Calculators - TI-108	06		________	________
Calculator Poster	01		________	________
Math Poster	01		________	________
Reading Poster	01		________	________
Tape Recorder/batteries	01		________	________

Section II. Miscellaneous

Sealing Tape	1	per box
Return Postage Paid Labels	1	per box
Ruler	10	
Geometric Shapes	10	
Post-it note pad	2	
Pad of Appointment Cards (40)	2	
Parent Information Letter	1	
Assessment Notification Letter	2	
Roster of Questionnaires	2	
Supplemental Shipping Envelope	1	
"Do Not Disturb" Sign	2	
Excluded Student Questionnaires	10	
School Characteristics and Policies Questionnaire (SCPQ)	1	
Teacher Questionnaires	10	
Cardboard	1	
Identification Sheet	2	
Bundle Slips	2	

Ship to: Assessment Administrator Name
Sherwood Elementary School
123 Main Street
Hometown, WI 12345

NAEP School #: 55A-116
Sherwood Elementary School

Session Type: Math Spiral
Reading Spiral

Assessment Date: 02/05/92

Makeup Date: ____________

Number of students to attend: ________

Packing Diagram

Packing List	top
Small Box containing: Calculators Cassette Tape Tape Recorder	Math Session only
Timer	
Roster of Questionnaires Completed Questionnaires	
Administration Schedule NAEP Identification Sheet Used Booklets	Band Booklets with Administration Schedule by Identification Sheet
Cardboard	
Posters Unused Questionnaires Unused Booklets	bottom

PLEASE RETURN ALL UNUSED MATERIALS

Scores for these items were gridded by the readers on separate, scannable scoring sheets, one sheet per booklet. As batches of test booklets cleared the editing process, scoring sheets for each batch of booklets were automatically generated by the system. Since the system had already captured all scannable information from each test booklet, scoring sheets could be generated for only those student booklets for which the student was present and eligible for the assessment. At the same time that the full set of scoring sheets was generated, the system randomly selected a 25 percent sample of booklets to be used for reliability scoring. A separate set of scoring sheets was generated for these booklets.

Each batch of scoring sheets was matched with the corresponding batch of student assessment booklets, and then forwarded to the professional scoring area. The scoring of the Trial State Assessment in reading was conducted concurrently with the scoring of the national assessment and the same readers scored the constructed-response items for both programs.

5.7.1 Description of Scoring

Each constructed-response item had a unique scoring standard that identified the range of possible scores for the item and defined the criteria to be used in evaluating the students' responses. The 60 readers scoring these items were organized into four teams of 15 readers, with one team leader per team. Each reader scored responses to 35 discrete short constructed-response items and 8 discrete extended constructed-response items at the fourth grade. The short constructed-response items were scored using a dichotomous scale of acceptable versus unacceptable response. The extended constructed-response items were scored using a graduated four-point scale:

1 = unsatisfactory response;
2 = partial response;
3 = essential response;
4 = extensive response.

Figure 5-4 shows the scoring guide used for one of the extended constructed-response reading items.

5.7.2 Training

The readers were trained by Educational Testing Service test development specialists to ensure that the teams would reliably score the constructed-response items. The training, conducted during a one-week period, served to familiarize the teams with the scoring standards. Scored sample papers were used to illustrate score point categorizations.

Before training began, the team leaders worked with ETS test development specialists to prepare training materials. Training consisted of first having the readers read each cognitive block passage; then the ETS trainer explained each item and its scoring rationale. The trainer then discussed sample responses that were representative of the various score points in the guide. Following the discussion of the scoring standards and the illustrative sample responses, the readers scored and discussed 35 to 50 "practice papers" for each extended constructed-

Figure 5-4
Extended Constructed-response Scoring Guide for "Sybil Sounds the Alarm"

Question

What are the major events in the story [Sybil Sounds the Alarm]?

Stance

Initial Understanding

General Scoring Rubric

Demonstrates an understanding of an historical narrative by summarizing the important major events.

1 = **Unsatisfactory** - These responses demonstrate little or no understanding of the events surrounding Sybil's ride by providing bits of information from the story, but not major events. In addition, these responses include those in which students merely copy one or more lines from the text, often the first or last sentence of the story.

2 = **Partial** - These responses demonstrate some understanding of Sybil's ride by providing an account of one or two major events, not usually accompanied a detailed account or an explanation of the importance of the events. These responses may also be a brief statement without specific events.

3 = **Essential** - These responses demonstrate an understanding of at least two of the major events surrounding Sybil's ride by providing a detailed account of these events **OR** by explaining the importance of the major events.

4 = **Extensive** - These responses demonstrate an in-depth understanding of the major events surrounding Sybil's ride by providing a detailed account of major events accompanied by an explanation of their significance. The responses display a thorough understanding of the story as a whole.

0 = **No response** (blank)

9 = **Not rateable** (I don't know, Off task, Illegible, etc.)

response item. During this practice, discussion sessions were held subsequent to the scoring of each 10 to 15 papers to review the scores assigned by the readers. Once the training on the extended constructed-response items was concluded, 10 complete sets of short constructed-response items were scored and discussed by the team.

Upon completing the practice sessions, the formal scoring process began. During the scoring, notes on various responses were compiled by the team leaders for the readers' reference and guidance. In addition, the team leaders met regularly to discuss particular responses; short training sessions were conducted when the team leaders determined that certain items were causing difficulty for the readers.

The team leaders conducted constant "back-reading" of all team members' work throughout the scoring process. The team leaders reviewed a percentage of the responses scored by each reader and brought any problems related to scoring to the attention of the individual reader. In this way, each team leader could be certain that his or her team was scoring consistently. When a reader's score was judged to be discrepant with the scoring standards, the team leader discussed the response and its score with that reader.

Upon completion of the 1992 constructed response scoring effort, it was determined that the interrater reliabilities (exact agreement) were not within the optimal range. Overall interrater reliability for the extended constructed-response questions was approximately 73 percent, whereas, the average reliability for regular constructed-response questions was about 89 percent. Several reasons were posited for the lower reliability of extended-responses. First, the purpose of extended constructed-responses questions is to tap more thoughtful and in-depth understandings that naturally require the use of more complex scoring rubrics. In addition, the number of scoring guides that scorers needed to be familiar with in order to accomplish the scoring of all blocks at one grade level may have created an overload of scoring standards, making it difficult for scorers to apply rubrics consistently across all items and responses. At the twelfth grade for example, there were a total of 86 constructed response scoring guides for professional scorers to learn. Also, training on all the scoring guides took place during a one week period before any scoring took place. The interspersing of regular constructed-response questions with extended constructed-response questions throughout the training and scoring process may have been somewhat problematic, as well, given the different types of scales involved. For the 1992 reading assessment, regular constructed-response questions were scored with a 2-point scoring guide, requiring scorers to make only acceptable versus unacceptable distinctions. However, extended constructed-response questions were rated in terms of four levels of comprehension, necessitating a more careful analysis of responses on the part of the scorer. These considerations were taken into account as a second scoring of extended constructed-response questions was planned after the initial scoring effort.

To determine whether or not the level of scorer agreement would be affected by having groups of scorers focus on responses to a single extended constructed-response question at a time, a team of 10 scorers and one team leader was selected to be trained and to score two of the eight extended constructed-response items with an item-by-item training/scoring procedure. This training session was conducted by the ETS test development specialist who had conducted the original training session. The original scoring standards were used. Immediately after the trial training session, a sample of approximately 120 papers that had been jointly scored by the ETS test development specialist and the team leader for each item were distributed to the readers as a means of gauging the degree of reader agreement on the score points. The outcome of this trial was that the percentages of reader agreement with these papers for these two items was approximately 90 percent. As a result of this successful trial, a second scoring session was conducted for all the extended constructed-response questions following the procedure used during the trial—scorers were trained on one scoring guide at a time and scoring for a single item took place immediately after training for that item.

This second scoring effort for extended constructed-response questions resulted in acceptable interrater reliabilities. These data are represented in Table 5-1 for the Trial State Assessment.

Table 5-1
Interreader Reliabilities for Extended Constructed-response Items
in the 1992 Trial State Assessment in Reading

NAEP ID	Description	Content Area	Interreader Reliability (Percent Exact Agreement)
R012006	Spider and Turtle	Literary Experience	88%
R012111	Box in the Barn	Literary Experience	93%
R012204	Blue Crabs	Gain Information	89%
R012305	Amanda Clements	Gain Information	85%
R012401	Sybil Sounds the Alarm	Literary Experience	90%
R012512	Watch Out for Wombats	Gain Information	91%
R012607	Money Makes Cares	Literary Experience	93%
R012708	Ellis Island	Gain Information	94%

5.7.3 Reliability of Scoring

Twenty-five percent of the booklets containing constructed-response items were scored by a second reader to obtain statistics on interreader reliability (see Table 5-1). At least 4,650 items were read twice. The average reliability for the 35 short constructed-response items was 88.54 percent. The average reliability for the eight extended constructed-response items was 73.55 percent during the first session of scoring and 90.37 percent during the second session of scoring. This reliability information was used by the team leaders in monitoring the accuracy of all individual readers and the uniformity of scoring across readers. Because the reliability scoring was done on separate scoring sheets, all reliability scoring was "blind," or uninfluenced by any score already assigned.

5.8 DATA TRANSCRIPTION SYSTEMS

The transcription of the student response data into machine-readable form was achieved through the use of three separate systems: data entry (scanning), validation (pre-edit), and resolution.

5.8.1 Data Entry

Machine-scannable booklets were used to collect the student response data from the 1992 reading assessment. These data were entered into the computer system using NCS optical scanning equipment. The data were then edited and questionable data were resolved before further processing.

To ensure data integrity, edit rules were applied to each scanned data field. This procedure validated each field and reported all problems for subsequent resolution. After each field was examined and corrected, the edit rules were re-applied for final verification.

5.8.2 Scanning

After the initial manual verification, the scannable documents were transported to a slitting area where the folded and stapled spine was removed from each document. Scanning operations were performed by NCS's HPS Optical Scanning equipment. The optical scanning devices and software used at NCS permits a complete mix of NAEP scannable materials to be scanned with no special grouping requirements. However, for manageability and tracking purposes, student documents, excluded student questionnaires, and teacher questionnaires were batched separately. In addition to the capture of scannable responses, the bar code identification numbers used to maintain process control were also decoded and transcribed to the NAEP computerized data file.

The scanning program is a table-driven software process that uses standard routines and application-specific tables to identify and define the documents and formats to be processed. When a booklet cover is scanned, the program uses the booklet number to determine the sequence of pages and the formats to be processed. By reading the booklet cover, the program recognizes which pages should follow and in what order.

The scanning program wrote four types of data records into the data set: a batch header record containing information coded onto the batch header sheet by receipt processing staff; a session header record containing information coded onto the session batch header sheet by receipt processing staff; a data record containing all of the translated marked ovals from all pages in a booklet; and a dummy data record, serving as a place holder in the file for a booklet with an unreadable cover sheet. The document code was written in the same location on all records to distinguish them by type.

The following coding rules were used:

- The data values from the booklet covers and scorer identification fields were coded as numeric data.
- Unmarked fields were coded as blanks and processing staff were alerted to missing or uncoded critical data.
- Fields that had multiple marks were coded as asterisks (*).
- The data values for the item responses and scores were returned as numeric codes.
- The multiple-choice, single-response format items were assigned codes depending on the position of the response alternative; that is, the first choice was assigned a 1, the second a 2, and so forth.
- The circle-all-that-apply items were given as many data fields as response alternatives; the marked choices were coded as 1 and the unmarked choices as blanks.
- The fields from unreadable pages were coded with an X as a flag for resolution staff to correct.

5.9 DATA VALIDATION

The data entry and resolution system used for the Trial State Assessment Program was also used for the national assessment program. The system is able to process materials submitted from both scannable and nonscannable media simultaneously for three age groups, three assessment types, and five questionnaires. The use of batch identification codes—comprising the school and session codes as well as the batch sequence numbers for suspect record identification—facilitated the management of the system and correction of incorrectly gridded or keyed information.

As the program processed each data record, it first read the booklet number and checked it against the batch session code for appropriate session type. Any mismatch was recorded on the error log and processing continued. The booklet number was compared against the first two digits of the student identification number. If they disagreed, because of improper bar coding, a message was written to the error log. The remaining booklet cover fields were then read and validated for the correct range of values. The school codes had to be identical to those on the Process Control System record and the grade code had to be 4. All data values that were out of range were read as is, but flagged as suspect. All data fields that were read as asterisks were recorded on the edit log.

Document definition files describe each document as a series of blocks that are described as a series of items. The blocks in a document were traversed in the order that they appear on the document. Each block's fields were validated during this process. If a document contained suspect fields, the cover information was recorded on the edit log with a description of the suspect data. Some fields (e.g., AGE or DOB), required special types of edits. These fields were identified in the document definition fields, and a subroutine was invoked to handle these cases.

The program next cycled through the data area corresponding to the item blocks. The task of translating, validating, and reporting errors for each data field in each block was performed by a routine that required only the block identification code and the string of input data. This routine had access to a block definition file that had the number of fields to be processed for each block and the field type (alphabetic or numeric), the field width in the data record, and the valid range of values for each field. The routine processed each field in sequential order, performing the necessary translation, validation, and reporting tasks.

The first of these tasks checked for the presence of blanks or asterisks in a critical field. These were recorded on the edit log and processing continued with the next field. No action was taken on blank-filled fields for multiple-choice items since that code indicated a nonresponse. The field was validated for range of response, recording anything outside of that range to the edit log. The item type code was used by the program to make a further distinction among constructed-response item scores and other numeric data fields. The last task performed in this processing phase was moving the translated and edited data field into the output buffer.

The completed string of data was written to the data file when the entire document had been processed. Then, when the next session header record was encountered, the program repeated the same set of processes for that session. The program closed the data set and generated an edit listing when it encountered the end of a file.

Accuracy checks were performed on each batch processed. Every 500th document of each booklet form was printed in its entirety, with a minimum of one document type per batch. This record was checked, item by item, with the source document for errors.

5.10 EDITING

Quality procedures and software throughout the system ensure that the NAEP data are correct. The initial editing that took place during the receipt control process included verification of the schools and sessions. Receipt control personnel checked that all student documents on the Administration Schedule were undamaged and assembled correctly. The machine edits performed during data capture verified that each sheet of each document was present and that each field had an appropriate value. All batches entered into the system were edited for errors.

Data editing occurred after these checks and consisted of a computerized edit review of each respondent's document and the clerical edits necessary to make corrections based upon the computer edit. This data editing step was repeated until all data were correct.

The first phase of data editing was designed to ensure that all documents were present. A computerized edit list was produced after NAEP documents were scanned and with the supporting documentation sent from the field the edit function was performed. The hard copy edit list contained all the vital statistics about the batch and each school and session within the batch, such as the number of students, school code, type of document, assessment code, error rates, suspect cases, and record serial numbers. Using these inputs, the data editor verified that the batch had been assembled correctly, each school number was correct, and all student documents within each session were present.

During data entry, counts of documents processed by type were generated. These counts were checked against the Administration Schedule counts entered into the Process Control System during the receiving process. The number of assessed and absent students processed had to match the number of used booklets indicated on the Process Control System.

The second phase of data editing was carried out by an experienced editing staff using a predetermined set of rules to review the field errors and record corrections to be made to the student data file. The same computerized edit list used in the first phase was also used to perform this function.

The editing staff made corrections using the edit log prepared by the computer and the actual source document listed on the edit log. The corrections were identified by batch sequence numbers and field name for suspect record and field identification. The edit log indicated the current composition of the field. This particular piece of information was then visually checked against the NAEP source document by the editing staff for double grids, erasures, smudge marks, or omitted items that were flagged. Each flagged item was handled in one of the following ways:

- *Correctable Error:* If the error could be corrected by the editing staff, according to the editing specifications, the corrections were indicated on the edit listing.

- *Field Correctable:* If an error was not correctable according to the specifications, an alert was issued to the operations coordinator for resolution. Once the correct information was obtained, the correction was indicated on the edit listing.

- *Noncorrectable Error:* If a suspected error was found to be correct as stated, and no alteration was possible according to source documents and specifications, the programs were tailored to allow this information to be accepted into the data record and no corrective action was taken.

These corrections were noted on the edit list. When the entire batch of sessions was resolved, the list was forwarded to the key entry staff. The corrections were entered and verified through the Falcon system. When all corrections were entered and verified for a batch, an extract program was run to pull the correction records to a mainframe data set.

The post-edit program was initiated next. This program applied the corrections to the specified records and once again applied the error criteria to all records. If there were further errors, another edit list was printed and the cycle began again.

When the edit process had produced an error-free file, the booklet ID number was posted to the NAEP tracking file by school and sessions. This allowed for an accumulation process to accurately measure the number of documents processed for a session within a school and the number of documents processed by form. The posting of booklet IDs also ensured that a booklet ID was not processed more than once. These data allowed the progress of the assessment to be monitored and reported on the status report.

At this point, a job was automatically submitted to produce the NAEP scoring sheets for this batch. The program also selected the records to be scored by a second reader for reliability. These sheets were printed, matched with the original documents, and forwarded to the NAEP scoring area.

Once all documents for a batch had been scored, the sheets were batched and submitted to scanning. A series of edits were run to verify the information on these sheets. The scorer identification fields were processed at this point and certain checks were made. The routine validated the score range and did not permit a blank field. If no score was indicated or the score was out of range, the disparity was noted on the edit log.

These error logs were returned to the scoring groups for resolution and the corrections were entered directly to the files. The edit process was repeated until the file was error free.

As a final quality control check, ETS identified a random sample of each booklet type from the master student file. The designated documents and scoring sheets were located, removed from storage and forwarded to ETS for quality control (see Chapter 6). On completion of quality control processing, the booklets were returned to NCS for return to storage.

5.11 QUESTIONNAIRES

The questionnaires were received either with the session shipment or in a later shipment. The questionnaires were checked against the roster and accumulated by the receiving clerks. The school characteristics and policies questionnaires, teacher questionnaires, and excluded student questionnaires were batched and sent to scanning at regular intervals. Every effort was made to keep current on all forms, both to ensure the processing of all documents for a session and to deliver all data at the same time.

All documents, regardless of method of entry, were run through the process of error identification and resolution.

5.12 MERGING OF STUDENT DATA

When the scoring and verification of the constructed responses was finished, the complete records for students were merged. This merge included the machine-scanned data and the scores for the constructed responses. Verification of complete student records was conducted prior to the delivery of the data files.

5.13 STORAGE OF DOCUMENTS

Once the editing process had been successfully completed on the batches, they were sent to the NCS warehouse for storage. The storage location of all documents was recorded on the inventory control system and stored for later retrieval. Unused materials were sent to temporary storage until the completion of the assessment and acceptance of the data files, at which time they were destroyed.

Chapter 6

CREATION OF THE DATABASE AND EVALUATION OF THE QUALITY CONTROL OF DATA ENTRY

John J. Ferris and David S. Freund

Educational Testing Service

6.1 OVERVIEW

The data transcription and editing procedures described in Chapter 5 resulted in the generation of disk and tape files containing various data for assessed students, excluded students, teachers, and schools. The weighting procedures described in Chapter 7 resulted in the generation of data files that included the sampling weights required to make valid statistical inferences about the population from which the 1992 fourth-grade Trial State Reading Assessment samples were drawn. These files were merged into a comprehensive, integrated database. To evaluate the effectiveness of the quality control of the data entry process, the final integrated database was sampled, and the data were verified in detail against the original instruments received from the field.

This chapter begins with a description of the transcribed data files and the procedure of merging them to create the 1992 Trial State Reading Assessment database for fourth-grade students. The last section presents the results of the quality control evaluation.

6.2 MERGING FILES INTO THE TRIAL STATE ASSESSMENT DATABASE

The transcription process conducted by National Computer Systems resulted in the transmittal to ETS of four data files for fourth grade: one file for each of the three questionnaires (teacher, school, and excluded student) and one file for the student response data. The sampling weights, derived by Westat, Inc., comprised an additional three files—one for students, one for schools, and one for excluded students. (See Chapter 7 for a discussion of the sampling weights.) These seven files were the foundation for the analysis of the 1992 Trial State Assessment data. Before data analyses could be performed, these data files had to be integrated into a coherent and comprehensive database.

The 1992 Trial State Reading Assessment database for fourth grade consisted of three files—student, school, and excluded student. Each record on the student file contained a student's responses to the particular assessment booklet the student was administered (booklets 30 to 45) and the information from the questionnaire that the student's reading teacher completed. (See Chapter 2 for information regarding assessment instruments.) Since teacher response data can be reported only at the student level, it was not necessary to have separate teacher files. The school files and excluded student files were separate and could be linked to the student files through the state and school codes.

The creation of the student data files began with the reorganization of the data files received from National Computer Systems. This involved two major tasks: 1) the files were restructured, eliminating unused (blank) areas to reduce the size of the files; and 2) in cases where students had chosen not to respond to an item, the missing responses were recoded as either "omitted" or "not reached," as appropriate. Next, the student response data were merged with the student weights file. The resulting file was then merged with the teacher response data. In both merging steps, the booklet ID (the two-digit booklet number and a five-digit serial number) was used as the matching criterion.

The school file was created by merging the school questionnaire file with the school weights file and a file of school variables, supplied by Westat, that included demographic information about the schools collected from the principal's questionnaire. The state and school codes were used as the matching criteria. Since some schools did not return a questionnaire and/or were missing principal's questionnaire data, some of the records in the school file contained only school-identifying information and sampling weight information.

The excluded student file was created by merging the excluded student questionnaire file with the excluded student weights file. The assessment booklet serial number was used as the matching criterion.

When the student, school, and excluded student files had been created, the database was ready for analysis. In addition, whenever new data values, such as composite background variables or plausible values, were derived, they were added to the appropriate database files using the same matching procedures as described above.

For archiving purposes, restricted-use data files and codebooks for each state were generated from this database. The restricted-use data files contain all responses and response-related data from the assessment, including responses from the student booklets and teacher and school questionnaires, proficiency scores, sampling weights, and variables used to compute standard errors.

6.3 CREATING THE MASTER CATALOG

A critical part of any database is its processing control and descriptive information. Having a central repository of this information, which may be accessed by all analysis and reporting programs, will provide correct parameters for processing the data fields and consistent labeling for identifying the results of the analyses. The Trial State Assessment master catalog file was designed and constructed to serve these purposes for the Trial State Assessment database.

Each record of the master catalog contains the processing, labeling, classification, and location information for a data field in the Trial State Assessment database. The control parameters are used by the access routines in the analysis programs to define the manner in which the data values are to be transformed and processed.

Each data field has a 50-character label in the master catalog describing the contents of the field and, where applicable, the source of the field. The data fields with discrete or categorical values (e.g., multiple-choice items and professionally scored items, but not weight fields) have additional label fields in the catalog containing 8- and 20-character labels for those values.

The classification area of the master catalog record contains distinct fields corresponding to predefined classification categories (e.g., reading content area) for the data fields. For a particular classification field, a nonblank value indicates the code of the subcategory within the classification categories for the data field. This classification area permits the grouping of identically classified items or data fields by performing a selection process on one or more classification fields in the master catalog.

The master catalog file was constructed concurrently with the collection and transcription of the Trial State Assessment data so that it would be ready for use by analysis programs when the database was created. As new data fields were derived and added to the database, their corresponding descriptive and control information were entered into the master catalog. The machine-readable catalog files are available as part of the secondary-use data files package for for use in analyzing the data with programming languages other than SAS and SPSS-X (see the *NAEP 1992 Trial State Assessment in Reading Secondary-use Data Files User Guide*).

6.4 QUALITY CONTROL EVALUATION

The purpose of the data entry quality control procedure is to gauge the overall accuracy of the process that transforms responses into machine-readable data. The procedure involves examining the actual responses made in a random sample of booklets and comparing them with the responses recorded in the final database, which is used for analysis and reporting.

6.4.1 Student Data

Sixteen assessment booklets numbered 30 through 45 were administered as part of the Trial State Assessment in reading. Table 6-1 provides the numbers of each booklet for which data were scanned into data files. These numbers varied somewhat more than in the 1990 assessment, but a chi-square measure of the variation proved to be nonsignificant at the 95 percent confidence level.

The number of students assessed in each of the 44 participating jurisdictions varied also. Twenty-nine jurisdictions met or exceeded the target of 2,500 students and a few smaller jurisdictions fell several hundred short of the target. The average number of students assessed in reading in each jurisdiction was 2,514.

Table 6-1

Number of Reading Booklets Scanned and Selected for Quality Control Evaluation

Booklet Number	Total Booklets Scanned	Total Booklets Selected
30	6,926	17
31	6,887	16
32	6,813	15
33	6,744	15
34	6,777	13
35	6,788	17
36	6,800	22
37	6,885	15
38	6,980	16
39	7,009	14
40	7,078	17
41	7,017	20
42	7,052	16
43	6,990	18
44	6,936	16
45	6,918	16
Total	110,600	263

To simplify the selection of booklets for examination, a method was developed that involved selecting all occurrences of a specified booklet in a randomly selected "stack." A stack is a unit of collection containing anywhere from 11 to 105 booklets, but typically between 50 and 60 booklets, in an assortment related to the spiraling technique used to distribute the booklets. The selection method was designed to yield approximately the same number of each booklet but, due to the variability in the size and contents of the stacks, there was somewhat more variation in the numbers of booklets selected than in the 1990 assessment (see Table 6-1). However, all of the booklets were sampled in adequate numbers and the average rate of selection was about one out of 440, a selection rate comparable to that used in past assessments at both the state and national levels. The few errors found during this quality control examination did not cluster by booklet number, so there is no reason to believe that the variation in numbers of booklets selected had a significant effect on the estimates of overall error rate confidence limits reported below.

The quality control evaluation detected only three errors in these student booklet samples—two instances of multiple responses that were not identified as such by the scanner, and one instance of an erasure that was recorded instead of ignored. As usual, there was some indication that the error rate could be improved with further tuning of the scanner procedures, but the process as it stands can certainly be described as clean and reliable. A very large volume of data was scanned with consistently excellent results. The usual quality control analysis based on the binomial theorem permits the inferences described in Table 6-2.

Table 6-2
Inference from the Quality Control Evaluation of Grade 4 Data

Subsample	Entry Type	Different Booklets Sampled	Number of Booklets Sampled	Characters Sampled	Number of Errors	Observed Rate	99.8% Confidence Limit
Student	Scanned	16	263	15,794	3	.0002	.0008
Teacher	Scanned	1	75	7,050	8	.0011	.0028
School	Scanned	1	97	9,312	4	.0004	.0015
Excluded Student	Scanned	1	66	5,148	4	.0008	.0027

6.4.2 Teacher Questionnaires

A total of 15,076 questionnaires were collected from reading teachers. Questionnaires were sampled at the rate of 1 in 200, resulting in the selection of 75 questionnaires. The selected questionnaires contained a total of eight errors, usually involving the scanner's mistaking an erasure for a response, but occasionally involving the failure of the scanner to pick up a multiple response. In every case, the respondent's intention was clear to the human eye, but the scanner seemed unprepared to exercise the same judgment that a careful observer would. The result is an error rate for the teacher questionnaire data that is about four times as high as for the student data. One possible explanation for this is that teacher questionnaires are inherently more complex than student assessment booklets, which leads to a much higher rate of erasures and other errors by the respondents. Perhaps a redesign of these questionnaires would bring the error rate down. This is not to say that the degree of erroneous data in the teacher questionnaire file is worrisome, but rather that the student data are more error-free. There is every indication that the quality of the teacher data is more than adequate for the purposes to which it was put.

6.4.3 School Questionnaires

A total of 4,857 questionnaires were collected from school administrators. These questionnaires were sampled for quality control evaluation at the rate of 1 in 50, resulting in the selection of 97 questionnaires. The quality of the data was very good, with an error rate of about half that of the teacher questionnaire data.

6.4.4 Excluded Student Questionnaires

A total of 13,268 excluded student questionnaires were scanned. These were sampled at the rate of about 1 in 200, resulting in the selection of 66 questionnaires. All the errors found were due to the scanner's mistaking an erasure for an intended response.

The quality of these data appears to be about as high as the other questionnaires—that is to say, adequate for the purposes to which it was put. The results of the evaluation of the questionnaire data are summarized in Table 6-2.

Chapter 7

WEIGHTING PROCEDURES AND VARIANCE ESTIMATION

Adam Chu and Keith F. Rust

Westat, Inc.

7.1 INTRODUCTION

Following the collection of assessment and background data from and about assessed and excluded students, sampling weights and associated sets of replicate weights were derived. The sampling weights are needed to make valid inferences from the student samples to the respective populations from which they were drawn. Replicate weights are used in the estimation of sampling variance, through the procedure known as jackknife repeated replication.

Each student was assigned a weight to be used for making inferences about the state's students. This weight is known as the *full* or *overall* sample weight. In the 1990 Trial State Assessment Program, a second weight, known as the comparison weight, was also derived for the purpose of comparing the assessment performance of students in monitored sessions with those in unmonitored sessions. However, for the 1992 Trial State Assessment Program, comparison weights were not calculated. Valid (i.e., unbiased) comparisons of this kind can be made using the full sample weights; however, the standard errors associated with these comparisons are somewhat larger than those that would be obtained using comparison weights.

The full-sample weight contains three components. First a base weight is established that is the inverse of the overall probability of selection of the sampled student. The base weight incorporates the probability of selecting a school and the student within a school, and accounts for the impact of procedures used to keep to a minimum the overlap of the state school sample with the NAEP national sample and the sample for the National Longitudinal Study of Chapter 1 Children (see Chapter 3). The base weight is then adjusted for two sources of nonparticipation—school-level and student-level. These weighting adjustments seek to reduce the potential for bias from such nonparticipation by increasing the weights of students from schools similar to those schools not participating, and increasing the weights of students similar to those students from within participating schools who did not attend the assessment session (or a makeup session) as scheduled. The details of how these weighting steps were implemented are given in sections 7.2 and 7.3.

In addition to the full-sample estimation weights, a set of replicate weights was provided for each student. These replicate weights are used in calculating the sampling errors of estimates obtained from the data, using the jackknife repeated replication method. Full details of the method of using these replicate weights to estimate sampling errors are contained in the technical reports for the 1988 and 1990 national assessments (Johnson & Zwick, 1990; Johnson & Allen, 1992). Section 7.5 of this report describes how the sets of replicate weights were generated for the 1992 Trial State Assessment data. The methods of deriving these weights were aimed at reflecting the features of the sample design appropriately in each state, so that when the jackknife variance estimation procedure is implemented, approximately unbiased estimates of sampling variance result.

7.2 CALCULATION OF BASE WEIGHTS

The base weight assigned to a school was the reciprocal of the probability of selection of that school. The school base weight depended on the subject of assessment since some schools were so small that students were tested in only one subject in those schools. In general, the school base weight reflected the actual probability used to select the school from the frame, including the impact of avoiding schools selected for the NAEP national sample and the sample for the National Longitudinal Study of Chapter 1 Children (see Chapter 3).

The student base weight was obtained by multiplying the school base weight by the within-school student weight, where the within-school student weight reflected the probability of selecting students within the school for a particular assessment subject. Additional details about the weighting process are given in the sections below.

7.2.1 Calculation of School Base Weights

As described in section 3.4.5, schools were sometimes selected in clusters in order to avoid giving small schools an extremely low probability of selection. The weight for sample cluster c was computed as:

$$W_c^{clust} = \frac{E}{mE_c}$$

where

E_c = the enrollment in the given grade for the cth cluster in the state;

E = $\sum_{c=1}^{M} E_c$

= the state-wide enrollment in the given grade; and

m = the number of clusters selected from the state.

In general, the base weight for sample school i in a given state was computed as:

$$W_i^{sch} = W_{ci}^{clust} T_{ci}$$

where W_{ci}^{clust} is the base weight of the cluster containing school i and T_{ci} is a "thinning" factor that reflects the fact that small schools in the Cluster Type 2 states were subject to thinning (see section 3.5.3). The thinning factor T_{ci} was equal to the ratio of the sampling size measure of the largest school in the cluster to the size measure of the retained school.

Since all schools in Cluster Type 1 states were included in the sample with certainty (see section 3.5.2), they were assigned school base weights (W_i^{sch}) equal to 1.

7.2.2 Weighting New Schools

As described in Chapter 3, new schools were sampled from the updated sampling frame list from each district in a sample of districts. In a few states, the selection probabilities of some new schools were quite small, resulting in excessively large school base weights. Where the weighted contribution to the estimate of total enrollment of a new school exceeded three times the median contribution, the base weight for that school was adjusted downwards (trimmed) in order to reduce the impact of the extreme weights on the variance of the estimates. Base weights were trimmed for one new school in New Jersey, North Carolina, and Ohio. For these three schools, the trimmed school weight (which was then used in the subsequent calculation of nonresponse adjustments) was computed as:

$$W_i^{sch} = \frac{E_{max}}{E_i}$$

where E_i is the estimated grade enrollment of the new school, and E_{max} is the maximum allowable weighted contribution to the estimated total grade enrollment for the given state. The value of E_{max} was established so that the weighted contribution of the new school to the total weighted grade enrollment never exceeded about three times the median value of the distribution of weighted enrollment counts for the remaining schools in the sample.

This adjustment was made to avoid introducing substantial variability into the sample estimates, as a result of giving relatively very large weights to one or two schools, and thus the sampled students within them. Although this procedure technically introduces a bias in the estimates for these states, we judged that it would be trivial in comparison to the level of sampling variance. For a discussion of issues involved in trimming of survey weights, see Potter (1988) and Stokes (1990).

7.2.3 Treatment of Substitute and Double-session Substitute Schools

Schools that replaced a refusing school (i.e., substitute schools) were assigned the weight of the refusing school, unless the substitute school also refused. Schools conducting extra sessions that served as substitutes for a refusing school (i.e., double-session substitutes) in effect had two school weights. The students in the school who were assigned to the original session were given the school base weight of the participating school, while those students assigned to the extra session(s) were assigned the school base weight of the refusing school.

7.2.4 Calculation of Student Base Weights

Within the sampled schools, eligible students were assigned to sessions using the procedures described in sections 3.5.7 and 3.6. The within-school probability of selection for assessment in reading therefore depended on the number of grade-eligible students in the school and the number of students selected for the assessment (usually 30 for a given subject). The within-school weights for the substitute schools were further adjusted to compensate for differences in the sizes of the substitute and the originally sampled (replaced) schools. The within-school weight also reflected the fact that a small school could have been selected for one subject but not the other. Thus, in general, the within-school student weight for the *j*th student in school *i* was equal to:

$$W_{ij}^{within} = \frac{N_i}{n_i} K_{1i} K_{2i}$$

where

N_i = the number of grade-eligible students enrolled in the school as reported in the sampling worksheets; and

n_i = the number of students selected for the given subject.

The factors K_{1i} and K_{2i} in the formula for the within-school student weight generally apply to only a few schools in each state. The factor K_{1i} adjusts the count of grade-eligible students in a substitute school to be consistent with corresponding count of the originally sampled (replaced) school. Specifically, for substitute schools,

$$K_{1i} = \frac{E_i}{E_i^M}$$

E_i = the QED grade 4 enrollment of the originally sampled (replaced) school; and

E_i^M = the QED grade 4 enrollment of the substitute school.

For nonsubstitute schools, $K_{1i} = 1$.

The factor K_{2i} reflects the subsampling procedure used to select the subject in which students in small schools were to be assessed (section 3.5.7). For a given subject, K_{2i} is defined as follows:

$$K_{2i} = \begin{cases} 1 & \text{if the fourth-grade school was selected for both subjects;} \\ 2 & \text{if the fourth-grade school was selected only for the given subject} \\ 0 & \text{if the fourth-grade school was not selected for the given subject} \end{cases}$$

Note that if K_{2i} is 2 for mathematics (say), then K_{2i} is 0 for reading, and vice versa.

The overall student base weight for a student *j* selected for reading assessment in school *i* was then computed as:

$$W_{ij}^{base} = W_i^{sch} W_{ij}^{within} .$$

Checks were made on these student base weights to ensure that the value was always 1.0 or greater.

7.3 Adjustments for Nonresponse

The base weight for a student was adjusted by two factors: one to adjust for nonparticipating schools for which no substitute participated, and one to adjust for students who were invited to the assessment but did not appear in either the scheduled or makeup sessions.

7.3.1 Defining Initial School-level Nonresponse Adjustment Classes

School-level nonresponse adjustment classes were initially created based on the urbanization and minority strata used in sampling. In states and urbanization strata where minority stratification was not used, nonresponse classes were created based on median household income.

The procedure for creating income classes was as follows. First, three classes of schools were formed for each urbanization stratum so that (1) each class had approximately the same number of sample schools and (2) the classes were ranked from low to high income. This was done using only the schools in the sample (including new schools), sorting them by median income, and then dividing the schools into three groups with equal numbers of schools. In a few states (Cluster Type 3 states) only large schools (those with grade enrollment over 20) were used to form the income strata, although all schools were classified into either income or minority strata. In creating the nonresponse adjustment classes, urbanization was used as the primary variable and minority/income was used as the secondary variable.

The initial nonresponse adjustment classes can be established for each state by considering the definitions of the sampling strata used, summarized in Table 3-2 of Chapter 3.

As can be seen in this tables, the definition of the initial nonresponse adjustment classes varied from one state to another. For example, nine classes obtained by cross-classifying three levels of urbanization (central city, suburban, other) with three levels of minority status (low, medium, and high) were defined for Alabama, whereas for New York, the classes were defined by minority status within the central city and suburban strata, and by income classes within the rural stratum, giving a total of 13 classes.

7.3.2 Constructing the Final Nonresponse Adjustment Classes

The objective in forming the final nonresponse adjustment classes was to create as many classes as possible that were internally as homogeneous as possible, but such that the resulting nonresponse adjustment factors were not subject to large random variation. The procedures discussed below were established with the aim of meeting this objective.

The schools were sorted into the initial nonresponse classes described above and the following unweighted and weighted counts and ratios were produced for each class:

- total in-scope schools from the original sample (an in-scope school is one that has at least one eligible student enrolled);
- participating in-scope schools from the sample (both original and substitutes); and
- total in-scope schools from the original sample divided by participating in-scope schools from the sample.

The weights used in the calculations were the school base weights defined in section 7.2, multiplied by the QED grade enrollment for the school.

The following guidelines were established for reviewing these counts and ratios and determining what collapsing should be done. Within an initial nonresponse class, if the weighted ratio of in-scope schools to participating schools was less than 1.35, with at least six participating schools in the class, there was no need to collapse the particular cell. If any nonresponse class had fewer than 6 schools or a ratio greater than or equal to 1.35, it was collapsed with another class such that the new class met these conditions. The order of variables to be collapsed (from most desirable to least desirable) was income strata or minority strata, followed by urbanization strata. The exceptions occurred in cases where minority classes within an urbanization stratum varied considerably as to the relative sizes of the minority population. In such cases, we collapsed over urbanization first to keep the classes as homogeneous as possible with regard to race/ethnicity. In some cases, final classes were formed with ratios in excess of 1.35. This occurred in states with relatively high school nonresponse. In no case was a class formed with fewer than six schools.

The choices of 1.35 as a cutoff for the nonresponse adjustment and 6 as the minimum number of participants within a class were both motivated by the desire to balance two conflicting needs. These are described in the first paragraph of this section. These limits were chosen on the basis of practical experience, combined with the application of theory about the effects of nonresponse class size on the accuracy of survey estimates, in a manner appropriate for the levels of nonresponse encountered in the various states.

7.3.3 School Nonresponse Adjustment Factors

The school-level nonresponse adjustment factor for the *i*th school in the *h*th class was computed as:

$$F_h^{(1)} = \frac{\sum_{i \in C_h} W_{hi}^{sch} E_{hi}}{\sum_{i \in C_h} W_{hi}^{sch} E_{hi} \delta_{hi}}$$

where

C_h = the subset of school records in class *h*;

W_{hi}^{sch} = the base weight of the *i*th school in class *h*;

E_{hi} = the QED grade enrollment for the *i*th school in class *h*;

$$\delta_{hi} = \begin{cases} 1 & \text{if the } i\text{th school in adjustment class } h \text{ participated in the assessments; and} \\ 0 & \text{otherwise.} \end{cases}$$

In the calculation of the above nonresponse adjustment factors, a school was said to have participated if

- it was selected for the sample from the QED frame or from the lists of new schools provided by participating school districts, and student assessment data were obtained from the school;
- the school refused but was replaced by a regular substitute school and student assessment data were obtained from the substitute school (so that the substitute participated in place of the originally selected school); or
- the school refused but was replaced by a double-session substitute school and the double-session substitute provided student assessment data for both the original and substitute sessions (so that the substitute school conducted additional sessions to replace the originally selected school).

Both the numerator and denominator of the nonresponse adjustment factor contained only in-scope schools.

The nonresponse-adjusted weight for the *i*th school in class *h* was computed as:

$$W_{hi}^{adj} = F_h^{(1)} W_{hi}^{sch} \quad .$$

7.3.4 Student-level Nonresponse Adjustment Classes

The variables used to define initial classes for adjusting for student nonresponse were:

- the final school-level nonresponse adjustment classes described in section 7.3.2;
- the age class of the student; and
- the monitor status of the session the student attended.

Two age classes, "old" and "young," were defined. "Old" students were those born in September 1981 or earlier; "young" students were those born after September 1981. Students in the "old" class are to some extent outliers with regard to age among their cohort. Previous findings from NAEP have shown that students in the "old" group tend to have higher absentee rates and lower proficiency scores than do students in the "young" group.

In order to determine whether the initial nonresponse classes needed collapsing, we reviewed the unweighted and weighted counts of assessed and absent students in each initial cell. (Excluded students were processed separately, using essentially the same procedures developed for assessed students.) The weight used for each student was the student base weight, adjusted for school nonresponse ($W_{hij}^{(2)}$ in section 7.3.5). The following guidelines were established for collapsing the initial nonresponse cells when necessary. Any cell with fewer than 20 assessed students was collapsed regardless of the value of the adjustment factor. If a cell had between 20 and 30 assessed students and the ratio of the weighted count of invited students to the weighted count of assessed students was greater than 1.5, the cell was collapsed. If a cell had more than 30 assessed students and the ratio of the weighted count of invited students to the weighted count of assessed students was greater than 2.0, the cell was collapsed.

When necessary, the collapsing of the initial cells proceeded as follows: First, collapsing was done across monitor status within all other classes. If the resulting cell still needed to be collapsed, the collapsing across monitor status was undone, and new cells were formed by collapsing across minority/income class. If these new cells still needed to be collapsed, collapsing across monitor status was done, followed by collapsing by urbanization class and finally by age group, if necessary. Based on these guidelines, some collapsing was done for all states, usually over monitor status and particularly for "old" students.

7.3.5 Student Nonresponse Adjustments

As described above, the student-level nonresponse adjustments for the assessed students were made within classes defined by the final school-level nonresponse adjustment cells, monitor status of the school, and age group of the students. Let the *k*th final (collapsed) nonresponse class be denoted as A_k. The adjusted student base weight for the *j*th sample student in school *i* in class A_k was calculated as:

$$W_{kij}^{(2)} = W_{hi}^{adj} W_{ij}^{within} = W_{hij}^{base} F_h^{(1)}$$

where

W_{hi}^{adj} = the nonresponse-adjusted school weight for school *i* in school adjustment class *h*;

W_{ij}^{within} = the within-school weight for the *j*th student in school *i*;

W_{hij}^{base} = $W_{hi}^{sch} W_{ij}^{within}$

= the student base weight for student *j* in school *hi*.

Using the adjusted student base weights, the assessed student nonresponse adjustment was calculated within nonresponse adjustment class A_k as:

$$F_k^{(2)} = \frac{\sum_{j \in A_k} W_{kij}^{(2)}}{\sum_{j \in A_k} W_{kij}^{(2)} \delta_{kj}}$$

where

$$\delta_{kj} = \begin{cases} 1 & \text{if the } j\text{th student in adjustment class } k \text{ participated in the assessments; and} \\ 0 & \text{otherwise.} \end{cases}$$

For excluded students, the same basic procedures as described above for assessed students were used, except that the numerator and denominator contained excluded rather than assessed students, and monitor status and student age group were not used to form the adjustment classes. An excluded student was regarded as a nonrespondent if no completed excluded student questionnaire was received.

The final student weight for the *j*th student in class *k* was then computed as:

$$W_{kj}^{final} = F_k^{(2)} W_{kj}^{(2)}$$

Tables 7-1 and 7-2 summarize the final unweighted and weighted counts of assessed and excluded students for each state. Checks were made on the final student weight distributions and totals at the state and subgroup within state, to ensure that there were no unexpected weight outliers or unusual distributions.

7.4 Characteristics of Nonresponding Schools and Students

In the previous section procedures were described for adjusting the survey weights so as to reduce the potential bias of nonparticipation of sampled schools and students. To the extent that a nonresponding school or student is different from those respondents in the same nonresponse adjustment class, potential for nonresponse bias remains.

In this section, we examine the potential for remaining nonresponse bias in two, related, ways. First we examine the weighted distributions, within each grade and state, of certain characteristics of schools and students, both for the full sample and for respondents only. This analysis is of necessity limited to those characteristics that are known for both respondents and nonrespondents, and hence cannot directly address the question of nonresponse bias. The approach taken does reflect the reduction in bias obtained through the use of nonresponse weighting adjustments. As such, it is more appropriate than a simple comparison of the characteristics of nonrespondents with those of nonrespondents for each state.

The second approach is to present some summary characteristics of nonrespondents and respondents from nonresponse adjustment classes where relatively large adjustment factors were obtained. In such classes the number of nonrespondents is relatively large, particularly in relation to the number of respondents available, and hence it is in these cases that the greatest potential for nonresponse bias exists. For those states and classes not appearing in these tables, it can be assumed that the potential for nonresponse bias is likely to be much less than in the cases shown.

7.4.1 Weighted Distributions of Schools Before and After School Nonresponse

Table 7-3 shows the mean values of certain school characteristics, both before and after nonresponse. The means are weighted appropriately to reflect whether nonresponse adjustments have been applied (i.e., to respondents only) or not (to the full set of in-scope schools). The variables for which means are presented are the percentage of students in the school who are Black, the percentage who are Hispanic, the median income of the ZIP code area where the school is located, and the "type of locale." All variables were obtained from the sample frame, described in Chapter 3. The type of locale variable has seven possible levels, which are defined in section 3.4.2. Although this variable is not interval-scaled, the mean value does give an indication of the degree of urbanization of the population represented by the school sample (lower values for type of locale indicate a greater degree of urbanization).

Table 7-1
Unweighted and Weighted Counts of Assessed Students by State

State	Grade 4/Reading	
	Unweighted	Weighted
Alabama	2,571	51,212
Arizona	2,677	48,310
Arkansas	2,589	32,074
California	2,365	323,231
Colorado	2,897	45,594
Connecticut	2,514	30,669
Delaware	2,048	66,162
District of Columbia	2,496	5,270
Florida	2,767	134,109
Georgia	2,712	89,643
Guam	2,029	2,154
Hawaii	2,642	12,718
Idaho	2,674	16,874
Indiana	2,535	70,397
Iowa	2,756	35,240
Kentucky	2,752	44,368
Louisiana	2,848	57,116
Maine	1,916	10,544
Maryland	2,786	54,036
Massachusetts	2,545	58,001
Michigan	2,437	111,584
Minnesota	2,589	54,335
Mississippi	2,657	36,892
Missouri	2,562	55,062
Nebraska	2,364	16,618
New Hampshire	2,239	13,927
New Jersey	2,239	74,747
New Mexico	2,305	20,970
New York	2,285	182,185
North Carolina	2,883	76,887
North Dakota	2,158	8,075
Ohio	2,580	132,772
Oklahoma	2,251	41,937
Pennsylvania	2,805	127,827
Rhode Island	2,347	10,037
South Carolina	2,758	47,615
Tennessee	2,734	57,700
Texas	2,571	243,738
Utah	2,829	34,607
Virgin Islands	882	1,823
Virginia	2,786	76,013
West Virginia	2,733	22,482
Wisconsin	2,712	57,983
Wyoming	2,775	7,867
TOTAL	110,600	2,701,405

Table 7-2
Unweighted and Weighted Counts of Excluded Students with Returned Questionnaires, by State

State	Grade 4/Reading	
	Unweighted	Weighted
Alabama	153	2,982
Arizona	213	3,737
Arkansas	151	1,813
California	399	54,701
Colorado	198	3,138
Connecticut	172	2,429
Delaware	137	456
District of Columbia	258	566
Florida	288	13,890
Georgia	154	5,169
Guam	153	154
Hawaii	166	833
Idaho	112	730
Indiana	113	3,182
Iowa	113	1,412
Kentucky	112	1,828
Louisiana	133	2,586
Maine	110	909
Maryland	190	3,771
Massachusetts	208	4,487
Michigan	133	5,793
Minnesota	104	2,352
Mississippi	149	1,985
Missouri	117	2,857
Nebraska	123	901
New Hampshire	112	637
New Jersey	133	4,510
New Mexico	185	1,740
New York	145	12,085
North Carolina	134	3,687
North Dakota	48	188
Ohio	174	8,846
Oklahoma	226	3,775
Pennsylvania	121	5,572
Rhode Island	183	763
South Carolina	168	2,927
Tennessee	135	3,158
Texas	250	20,879
Utah	140	1,630
Virgin Islands	32	66
Virginia	190	5,214
West Virginia	148	1,249
Wisconsin	196	4,235
Wyoming	123	337
TOTAL	7,002	204,159

Table 7-3
Weighted Mean Values Derived from Sampled Schools, Grade 4

State	Weighted Participation Rate After Substitution	Weighted Mean Values Derived from Full Sample				Weighted Mean Values Derived from Responding Sample, with Substitutes and School Nonresponse Adjustment			
		Percent Black	Percent Hispanic	Median Income	Type of Locale	Percent Black	Percent Hispanic	Median Income	Type of Locale
Alabama	76%	31.80	0.04	$22,374	4.66	31.54	0.04	$22,471	4.66
Arizona	99%	4.05	21.75	$29,744	3.18	4.08	21.80	$29,742	3.19
Arkansas	87%	24.31	0.40	$21,357	5.39	23.55	0.40	$21,415	5.42
California	92%	8.31	35.92	$32,603	3.18	8.33	35.88	$32,747	3.16
Colorado	100%	4.49	17.11	$31,493	3.67	4.49	17.11	$31,493	3.67
Connecticut	99%	9.88	8.52	$39,525	3.64	9.88	8.52	$39,555	3.63
Delaware	92%	24.25	0.32	$25,543	4.48	23.20	0.32	$25,290	4.48
Dist. of Columbia	99%	90.59	3.69	$27,879	1.00	90.58	3.70	$27,821	1.00
Florida	100%	24.10	10.94	$27,508	3.60	24.10	10.94	$27,508	3.60
Georgia	100%	33.93	1.34	$28,190	4.41	33.93	1.34	$28,190	4.41
Guam	100%	2.27	0.31	-	7.00	2.27	0.31	-	7.00
Hawaii	100%	1.41	0.00	$34,004	3.98	1.41	0.00	$34,004	3.98
Idaho	82%	0.12	4.78	$25,466	5.44	0.12	4.82	$25,501	5.43
Indiana	77%	11.44	0.59	$28,432	4.33	11.08	0.58	$28,538	4.35
Iowa	100%	0.95	0.25	$26,153	4.92	0.96	0.25	$26,153	4.92
Kentucky	94%	7.38	0.07	$22,637	5.27	7.39	0.06	$22,609	5.27
Louisiana	100%	44.82	0.82	$22,398	4.28	44.82	0.82	$22,398	4.28
Maine	58%	0.17	0.55	$27,037	5.71	0.08	0.53	$26,812	5.70
Maryland	99%	27.84	1.36	$39,703	3.46	27.72	1.38	$39,923	3.46
Massachusetts	87%	6.75	4.11	$37,162	3.70	6.77	4.11	$37,160	3.69
Michigan	83%	14.66	0.95	$31,737	4.12	14.41	1.10	$31,794	4.12
Minnesota	81%	2.05	0.54	$32,278	4.70	1.99	0.53	$32,529	4.71
Mississippi	98%	48.16	0.17	$19,464	5.56	48.16	0.17	$19,464	5.56
Missouri	90%	15.19	0.64	$27,091	4.50	15.06	0.63	$26,941	4.54
Nebraska	76%	3.95	0.94	$27,729	4.77	3.84	1.14	$27,709	4.78
New Hampshire	68%	0.74	0.85	$35,664	5.21	0.66	0.70	$35,635	5.21
New Jersey	76%	15.63	8.27	$40,407	3.60	14.73	8.56	$40,204	3.60
New Mexico	76%	2.61	44.28	$22,576	4.63	2.75	45.61	$22,810	4.63
New York	78%	15.59	15.87	$32,148	3.20	14.57	15.96	$32,263	3.21
North Carolina	95%	27.45	0.01	$26,040	4.95	27.12	0.01	$26,105	4.95
North Dakota	70%	0.47	0.06	$26,971	5.05	0.25	0.07	$26,890	5.06
Ohio	78%	10.31	0.35	$28,808	4.13	9.61	0.33	$28.970	4.16
Oklahoma	86%	7.29	1.26	$25,298	4.54	7.35	1.27	$25,318	4.54
Pennsylvania	85%	12.67	3.26	$28,430	4.29	12.83	3.28	$28,435	4.28
Rhode Island	83%	4.30	3.86	$30,172	3.37	3.90	4.01	$30,065	3.36
South Carolina	98%	37.52	0.07	$26,484	4.99	37.43	0.07	$26,519	4.99
Tennessee	93%	20.81	0.06	$24,438	4.14	21.12	0.06	$24,614	4.13
Texas	92%	14.47	34.28	$26,315	3.44	14.29	34.66	$26,281	3.43
Utah	99%	0.13	0.83	$31,112	4.25	0.13	0.84	$31,122	4.25
Virginia	99%	24.82	1.37	$36,554	4.19	24.79	1.37	$36,381	4.19
Virgin Islands	100%	82.99	15.29	-	9.00	82.99	15.29	-	-
West Virginia	100%	2.86	0.18	$21,639	5.60	2.86	0.18	$21,639	5.60
Wisconsin	99%	6.74	1.32	$31,270	4.39	6.71	1.32	$31,216	4.40
Wyoming	97%	0.63	6.53	$30,859	5.45	0.64	6.63	$30,806	5.45

Two sets of means are presented for these four variables. The first set shows the weighted mean derived from the full sample of in-scope schools selected for reading; that is, respondents and nonrespondents (for which there was no participating substitute). The weight for each sampled school is the product of the school base weight and the grade enrollment. This weight therefore represents the number of students in the state represented by the selected school. The second set of means is derived from responding schools only, after school substitution. In this case the weight for each school is the product of the nonresponse-adjusted school weight and the grade enrollment, and therefore indicates the number of students in the state represented by the responding school.

The differences between these sets of means give an indication of the potential for nonresponse bias that has been introduced by nonresponding schools for which there was no participating substitute. For example, in Arkansas at grade 4 the mean percentage Black enrollment, estimated from the original sample, is 24.31 percent. The estimate from the responding schools is 23.55 percent. Thus there may be a slight bias in the results for Arkansas because these two means differ. Note, however, that throughout these two tables the differences in the two sets of mean values are very slight, suggesting that it is unlikely that substantial bias has been introduced by schools that did not participate and for which no substitute participated. Of course in a number of states (as indicated) there was no nonresponse at the school level, so that these sets of means are identical. Even in those states where school nonresponse was relatively high (such as Maine, New Jersey, and New York), the differences in means are slight.

7.4.2 Characteristics of Nonresponding Schools

Table 7-4 shows the distributions of some characteristics of nonresponding and responding schools, by school nonresponse adjustment class, for classes with adjustment factors in excess of 1.25. The respondents include the case where substitute schools participated. In other words, the nonrespondents include only those nonrespondents for which no substitute participated.

The characteristics shown are as follows:

- *The set of distinct values for the "type of locale" variable.* This variable, which was used for sample stratification, has seven possible levels, which are defined in Chapter 3, section 3.4.2.
- *The percentage of the state's public-school fourth-grade enrollment represented in the sample by the schools within the adjustment class.* The school nonresponse adjustment factor is calculated directly from these two quantities (one for respondents, one for nonrespondents). The potential for nonresponse bias is generally greater in cases where the size of the set of nonrespondents is relatively large.
- *The minimum, median, and maximum percentage enrollments of Black and Hispanic students.* In cases where there are only two nonresponding school/hits involved, only the minimum and maximum are presented.

Table 7-4
Grade 4 School Nonresponse Adjustment Classes with Adjustment Factors Greater than 1.25

State	Class	Nonresponse Adjustment	Response Status	Number of School Selections	Types of Locale	Percent of State Student Population Represented	Enrollment Percent Black			Enrollment Percent Hispanic			Median Household Income ($)*		
							Min.	Med.	Max.	Min.	Med.	Max.	Min.	Med.	Max.
Delaware	10	1.30093	Respondents	18	7	26.08%	0%	20%	50%	0%	0%	0%	20,300	24,158	30,976
			Nonrespondents	6	7	7.85%	10%	40%	40%	0%	0%	0%	23,655	28,723	28,723
Massachusetts	7	1.28355	Respondents	10	5,6,7	7.27%	0%	0%	10%	0%	0%	0%	24,140	30,252	33,931
			Nonrespondents	3	6,7	2.07%	0%	0%	0%	0%	0%	0%	25,519	28,655	31,577
Maine	3	1.35928	Respondents	14	6	12.29%	0%	0%	0%	0%	0%	1%	20,263	23,062	24,506
			Nonrespondents	6	6	4.41%	0%	0%	8%	0%	0.5%	3%	21,226	22,229	24,498
	4	1.86382	Respondents	10	6	7.62%	0%	0%	3%	0%	0.5%	4%	24,573	25,674	28,452
			Nonrespondents	9	6	6.58%	0%	0%	1%	0%	1%	2%	25,747	27,924	28,531
	5	1.44558	Respondents	13	6	11.71%	0%	0%	1%	0%	1%	2%	28,552	30,947	40,726
			Nonrespondents	6	6	5.22%	0%	0%	1%	0%	0%	2%	28,912	31,775	50,455
	6	1.74923	Respondents	11	7	4.68%	0%	0%	0%	0%	0%	1%	14,106	19,504	22,585
			Nonrespondents	9	7	3.51%	0%	0%	0%	0%	0%	1%	14,879	20,735	22,206
	8	1.53416	Respondents	12	7	7.17%	0%	0%	0%	0%	0%	1%	26,236	28,550	38,719
			Nonrespondents	7	7	3.83%	0%	0%	0%	0%	0%	1%	26,507	27,251	45,690
Michigan	1	1.28032	Respondents	7	2	5.99%	0%	10%	20%	0%	0%	0%	26,546	28,607	40,794
			Nonrespondents	2	2	1.68%	10%	--	20%	0%	–	0%	22,972	–	30,675
	9	1.27040	Respondents	7	5,6	6.93%	0%	0%	0%	0%	0%	0%	35,411	37,834	44,411
			Nonrespondents	3	6	1.87%	0%	0%	0%	0%	0%	1%	32,740	34,043	38,501
Minnesota	1	1.26531	Respondents	10	1,2	9.80%	0%	0%	30%	0%	0%	1%	22,370	32,283	47,213
			Nonrespondents	3	1,2	2.60%	0%	0%	31%	0%	0%	2%	16,845	20,316	35,877
Nebraska	4	1.36279	Respondents	19	5,66	13.36%	0%	0%	0%	0%	0%	50%	18,814	25,837	27,647
			Nonrespondents	11	5,6	4.85%	0%	0%	0%	0%	0%	7%	21,910	25,188	27,724
	6	1.35796	Respondents	29	7	15.98%	0%	0%	0%	0%	0%	0%	14,310	20,486	23,436
			Nonrespondents	15	7	5.72%	0%	0%	0%	0%	0%	0%	8,332	19,636	23,476

---

* Median household income of ZIP code area where school is located, derived from 1980 population census data and expressed in 1985 dollars.

Table 7-4 (continued)
Grade 4 School Nonresponse Adjustment Classes with Adjustment Factors Greater than 1.25

State	Class	Nonresponse Adjustment	Response Status	Number of School Selections	Types of Locale	Percent of State Student Population Represented	Enrollment Percent Black			Enrollment Percent Hispanic			Median Household Income ($)*		
							Min.	Med.	Max.	Min.	Med.	Max.	Min.	Med.	Max.
New Hampshire	2	1.28341	Respondents	7	2,4	7.11%	0%	1%	3%	0%	0%	3%	33,697	36,315	38,374
			Nonrespondents	2	2	2.02%	3%	–	3%	3%	–	16%	33,067	–	33,067
	4	1.31487	Respondents	15	6	13.95%	0%	1%	2%	0%	0%	2%	21,480	27,669	29,618
			Nonrespondents	5	6	4.39%	0%	1%	1%	0%	0%	1%	25,044	28,463	28,524
	5	1.25814	Respondents	16	5,6	13.96%	0%	0.5%	4%	0%	0%	2%	31,172	33,963	39,464
			Nonrespondents	4	6	3.60%	0%	0.5%	1%	0%	0%	1%	30,971	33,061	36,831
	6	1.37357	Respondents	14	5,6	13.65%	0%	0%	1%	0%	0%	2%	39,464	44,991	56,506
			Nonrespondents	5	6	5.10%	0%	1%	1%	0%	1%	1%	41,444	46,202	63,873
New Jersey	5	1.25800	Respondents	38	3	31.82%	0%	0%	95%	0%	0%	50%	21,861	41,472	72,889
			Nonrespondents	8	3	8.21%	0%	0%	80%	0%	0%	10%	32,153	42,112	75,500
	6	1.43144	Respondents	16	4	13.46%	0%	0%	30%	0%	0%	72%	23,088	46,631	72,339
			Nonrespondents	8	4	5.81%	0%	0%	50%	0%	0%	0%	34,731	43,795	55,632
	8	1.31707	Respondents	7	5,6,7	5.22%	0%	0%	40%	0%	0%	2%	36,934	39,242	46,384
			Nonrespondents	2	6,7	1.65%	0%	–	0%	0%	–	0%	44,135	–	45,587
	9	1.26542	Respondents	6	5,6,7	5.93%	0%	0%	0%	0%	0%	0%	52,563	61,042	68,749
			Nonrespondents	2	6,7	1.58%	0%	–	0%	0%	–	0%	50,521	–	56,951
New Mexico	7	1.38274	Respondents	7	6	7.42%	0%	0%	17%	4%	11%	25%	15,417	23,665	44,495
			Nonrespondents	3	6	2.84%	0%	–	1%	0%	12%	31%	13,792	21,755	24,393
	10	1.30943	Respondents	19	7	12.42%	0%	0%	2%	0%	63%	99%	10,393	16,838	24,393
			Nonrespondents	6	7	3.84%	0%	0%	3%	0%	29%	73%	13,988	16,076	26,480
New York	2	1.35294	Respondents	17	1,2	16.18%	1%	21%	57%	25%	55%	97%	13,243	18,701	35,674
			Nonrespondents	6	1,2	5.71%	3%	35%	79%	20%	54%	91%	12,271	15,599	24,336
	3	1.32967	Respondents	9	1,2	8.66%	32%	45%	94%	0%	2%	13%	16,589	22,304	39,448
			Nonrespondents	3	1	2.86%	38%	92%	94%	5%	8%	17%	11,411	15,319	31,124

---

* Median household income of ZIP code area where school is located, derived from 1980 population census data and expressed in 1985 dollars.

Table 7-4 (continued)
Grade 4 School Nonresponse Adjustment Classes with Adjustment Factors Greater than 1.25

State	Class	Nonresponse Adjustment	Response Status	Number of School Selections	Types of Locale	Percent of State Student Population Represented	Enrollment Percent Black			Enrollment Percent Hispanic			Median Household Income ($)*		
							Min.	Med.	Max.	Min.	Med.	Max.	Min.	Med.	Max.
North Dakota	2	1.34178	Respondents	9	2,4	8.25%	0%	0%	2%	0%	0%	2%	30,678	30,678	31,202
			Nonrespondents	3	2	2.82%	0%	0%	0%	0%	0%	0%	30,678	31,202	32,980
Ohio	9	1.34462	Respondents	6	6	4.88%	0%	0%	10%	0%	0%	20%	27,323	29,847	36,663
			Nonrespondents	2	5,6	1.68%	0%	–	0%	0%	–	10%	28,550	–	28,972
	10	1.29348	Respondents	7	7	5.53%	0%	0%	0%	0%	0%	0%	13,471	21,213	23,580
			Nonrespondents	3	7	1.62%	0%	0%	0%	0%	0%	0%	17,135	21,982	23,078
Tennessee	8	1.26907	Respondents	9	5,6	7.20%	0%	0%	45%	0%	0%	1%	19,287	19,960	21,905
			Nonrespondents	2	6	1.94%	0%	–	0%	0%	–	0%	20,319	–	21,111

---

* Median household income of ZIP code area where school is located, derived from 1980 population census data and expressed in 1985 dollars.

- *The minimum, median, and maximum household incomes of the five digit ZIP code area where the school is located.* The data are calculated from 1980 Census data, but are updated to 1985 dollars. Note that the small numbers of nonresponding schools in each class, and the fact that data is at the ZIP code area level, means that on occasion the median and maximum values, for example, are identical.

Examination of the table shows that invariably the respondents and nonrespondents are quite similar with regard to type of locale. There are great similarities in many cases for other characteristics also, but on some occasions the nonresponding schools have a somewhat lower median income distribution than the respondents, and occasionally also there is some difference in the distributions of minority enrollment levels. For example, in New York, Class 3, the nonresponding schools have somewhat higher rates of Black and Hispanic enrollment and somewhat lower median household incomes than the respondents. By contrast, in New Mexico, Class 10, the nonresponding schools have somewhat lower Hispanic enrollment and noticeably higher median income than the respondents. In Minnesota, Class 1, the nonresponding schools have somewhat lower median income than the respondents.

7.4.3 Weighted Distributions of Students Before and After Student Absenteeism

Table 7-5 shows, for each state, the weighted sampled percentages of students by gender (male) and race/ethnicity (White, not Hispanic; Black, not Hispanic; Hispanic) for the full sample of students (after student exclusion) and for the assessed sample.

The weight used for the full sample is the adjusted student base weight, defined in section 7.3.5. The weight for the assessed students is the final student weight, also defined section 7.3.5. The difference between the estimates of the population subgroups is an estimate of the bias in estimating the size of the subgroup, resulting from student absenteeism from the assessment. As such it is an indicator of the potential for nonresponse bias in the assessment results, resulting from student absenteeism.

Care must be taken in interpreting these results, however. First, note that there is generally very little difference in the proportions estimated from the full sample and those estimated from the assessed students. While this is encouraging, it does not eliminate the possibility that bias exists, either within the state as a whole, or for results for gender and race/ethnicity subgroups, or for other subgroups. Second, on the other hand, where differences do exist they cannot be used to indicate the likely magnitude or direction of the bias with any reliability. For example, in New Jersey, the percentages of Black and Hispanic students in the full sample are respectively 15.65 and 13.85 percent. For assessed students, these percentages are 14.23 for Black students and 13.08 for Hispanic students. While these differences raise the possibility that some bias exists, it is not appropriate to speculate on the magnitude of this bias by considering the assessment results for Black and Hispanic students, in comparison to other students in the state. This is because the underrepresented Black and Hispanic students may not be typical of students that were included in the sample, and similarly those students within the same racial/ethnic groups who are disproportionately overrepresented may not be typical either. This is because not all students within the same race/ethnicity group receive the same student nonresponse adjustment. Some insight as to the kinds of students who are receiving

Table 7-5
Weighted Student Percentages Derived from Sampled Schools, Grade 4

State	Weighted Student Participation	Weighted Percentages Derived from Full Sample				Weighted Percentages Derived from Assessed Sample, with Student Nonresponse Adjustment			
		Percent Male	Percent White	Percent Black	Percent Hispanic	Percent Male	Percent White	Percent Black	Percent Hispanic
Alabama	96%	51.31	61.50	31.52	4.57	51.82	61.48	31.27	4.80
Arizona	95%	48.67	56.06	4.21	28.08	48.44	55.63	4.18	28.59
Arkansas	96%	50.07	69.69	20.69	6.47	50.15	69.53	20.50	6.74
California	94%	49.23	45.73	7.00	34.67	49.21	45.53	6.76	34.66
Colorado	95%	51.22	70.02	4.05	20.81	50.95	69.99	4.00	21.04
Connecticut	95%	51.17	72.77	11.02	13.07	50.71	72.91	10.79	13.08
Delaware	95%	50.27	63.84	24.85	7.71	50.26	63.69	24.73	8.04
Dist. of Columbia	94%	49.66	5.00	83.01	9.12	49.73	4.93	82.94	9.29
Florida	95%	40.33	57.52	21.01	17.65	50.57	57.00	21.28	17.78
Georgia	96%	50.75	57.56	34.06	5.20	50.86	57.49	33.90	5.43
Guam	94%	51.43	12.44	3.80	16.52	51.65	12.21	3.95	17.51
Hawaii	95%	50.87	19.97	4.78	18.36	50.62	19.65	5.03	19.31
Idaho	96%	49.78	84.41	0.59	10.45	49.58	83.99	0.58	10.76
Indiana	96%	49.90	81.63	11.12	5.06	49.76	81.62	10.84	5.34
Iowa	96%	49.89	88.46	2.95	5.42	49.74	88.15	3.07	5.57
Kentucky	96%	52.58	86.42	9.12	2.99	52.50	86.42	8.97	3.10
Louisiana	96%	50.49	51.41	41.25	4.54	50.02	51.21	41.15	4.70
Maine	95%	48.24	92.51	0.45	4.37	48.32	92.48	0.44	4.45
Maryland	95%	49.21	60.27	29.44	5.73	49.38	60.16	29.34	5.91
Massachusetts	96%	50.64	81.34	7.29	7.00	50.42	81.37	7.20	7.11
Michigan	94%	49.92	73.42	14.24	8.22	49.67	73.97	13.27	8.48
Minnesota	96%	51.37	87.31	3.03	5.40	51.46	87.08	3.10	5.68
Mississippi	97%	51.87	41.18	52.55	4.91	51.70	41.02	52.49	5.10
Missouri	95%	50.26	77.10	14.50	5.14	50.15	77.13	14.05	5.43
Nebraska	96%	51.73	82.95	6.46	7.83	52.23	83.50	5.72	8.03
New Hampshire	96%	50.48	90.31	0.84	4.89	50.77	90.22	0.81	4.96
New Jersey	96%	50.23	64.61	15.65	13.84	50.06	66.57	14.22	13.07
New Mexico	95%	50.08	45.11	3.37	45.77	49.81	45.13	2.92	46.01
New York	95%	51.57	63.05	12.77	18.56	51.73	60.54	13.63	20.00
North Carolina	96%	50.44	63.15	27.73	5.02	50.57	62.71	27.96	5.17
North Dakota	97%	50.42	93.05	0.45	2.96	50.67	92.87	0.47	3.06
Ohio	96%	49.20	80.50	12.24	5.06	49.52	81.01	11.58	5.28
Oklahoma	85%	49.25	72.81	8.34	7.28	48.81	72.35	7.83	8.15
Pennsylvania	95%	48.42	78.02	11.70	7.68	48.24	78.59	10.96	7.73
Rhode Island	95%	50.62	76.44	6.27	11.81	50.83	76.41	5.97	12.02
South Carolina	96%	48.33	55.04	37.82	4.98	48.08	54.81	37.79	5.17
Tennessee	95%	50.31	71.24	21.24	5.00	50.26	71.20	21.04	5.16
Texas	96%	51.70	47.77	13.76	35.05	51.61	48.61	13.53	34.40
Utah	96%	48.44	86.01	0.71	9.40	48.33	85.82	0.75	9.62
Virginia	96%	51.34	67.31	23.86	4.82	51.08	67.37	23.53	4.98
Virgin Islands	97%	52.15	2.52	75.86	19.41	51.93	2.38	75.88	19.56
West Virginia	96%	50.46	91.25	2.33	4.01	50.59	91.16	2.32	4.16
Wisconsin	96%	49.78	82.64	5.66	7.86	49.89	82.54	5.59	8.06
Wyoming	96%	50.60	83.02	0.61	11.30	50.95	82.75	0.60	11.59

relatively large adjustments, and the kinds of students that they are being adjusted to represent, are given in the next section. Small sample sizes within nonresponse adjustment classes make this information difficult to interpret, however. One other feature to note is that, for assessed students, information as to the student's gender and race/ethnicity is provided by the student, while for absent students this information is provided by the school. Evidence from past NAEP assessments (see, for example, Rust & Johnson, 1992) indicates that there can be substantial discrepancies between those two sources, especially with regard to classifying students as Hispanic at grade 4.

7.4.4 Characteristics of Absent Students

Table 7-6 shows some characteristics of assessed (responding) and absent (nonresponding) students, by student nonresponse adjustment class, for classes with adjustment factors in excess of 1.25.

In addition to information characterizing the class in terms of age class, monitor status, and type of location, the distributions of certain characteristics of assessed and absent students within each class are presented. The characteristics shown are:

- *The percentage of the state's public-school grade enrollment represented in the sample by the students within the adjustment class.* This is given by the sum of the adjusted student base weights ($W_{hij}^{(2)}$, see section 7.3.5) for the responding and nonresponding selected students respectively, within the student-level nonresponse adjustment class. The student nonresponse adjustment factor is calculated directly from these two quantities (one for respondents, one for nonrespondents). The potential for nonresponse bias is generally greater in cases where the size of the population represented by the nonrespondents is relatively large.

- *The percentage of students who are male, weighted by the base weight for each student adjusted for school nonresponse.* This estimates the proportion of students who are male in the subpopulation represented by the sample students.

- *The percentages of students who are White, Black, Hispanic, or of another race/ethnicity.* Again these percentages are weighted by the students' base weights, adjusted for school nonresponse.

The table shows that assessed and absent students have similar characteristics within nonresponse adjustment classes. A notable feature is that most of the cases involving adjustment factors in excess of 1.25 occur within classes in which the students are in age class 1—that is, relatively old for their grade. Since both the respondents and nonrespondents share this characteristic, this is not in itself a source of nonresponse bias. The potential for bias arises because of the possibility that, within this group, the respondents differ from the nonrespondents.

Table 7-6
Grade 4 Student Nonresponse Adjustment Classes with Adjustment Factors Greater than 1.25

State	Class	Nonresponse Adjustment	Age Class*	Monitor Status	Types of Locale	Response Status	Percent of State Population	Percent Male	Percent Race/Ethnicity			
									White	Black	Hispanic	Other
Michigan	2	1.25501	1	Both	1,2	Respondents	1.04%	58.6%	36.1%	37.1%	26.8%	0.0%
						Nonrespondents	0.26%	61.4%	0.0%	55.1%	44.9%	0.0%
Oklahoma	1	1.32301	1	Unmonitored	1,2	Respondents	2.79%	65.1%	69.2%	4.8%	10.6%	15.5%
						Nonrespondents	0.90%	57.1%	88.3%	0.0%	7.6%	4.1%
	3	1.28784	1	Unmonitored	1,2	Respondents	2.04%	42.0%	37.5%	50.4%	7.4%	4.6%
						Nonrespondents	0.59%	61.7%	50.0%	35.2%	9.8%	4.9%
	4	1.53102	1	Monitored	1,2	Respondents	1.44%	57.9%	56.0%	18.4%	18.2%	7.4%
						Nonrespondents	1.76%	55.9%	52.5%	27.2%	6.7%	13.6%
	7	1.25852	1	Both	3,4	Respondents	1.19%	60.4%	85.5%	2.7%	9.2%	2.6%
						Nonrespondents	0.31%	100%	90.0%	10.0%	0.0%	0.0%
	8	1.25812	1	Both	6	Respondents	1.01%	53.7%	68.4%	0.0%	6.6%	24.9%
						Nonrespondents	0.26%	59.2%	73.2%	14.0%	0.0%	12.8%
	11	1.25812	1	Monitored	6	Respondents	1.88%	51.8%	69.1%	6.9%	11.7%	12.3%
						Nonrespondents	0.80%	73.7%	75.1%	9.6%	0.0%	15.3%
	12	1.33498	1	Unmonitored	5,6	Respondents	2.63%	62.8%	83.1%	1.3%	8.9%	6.8%
						Nonrespondents	0.88%	75.3%	85.9%	0.0%	4.5%	9.6%
	27	1.34143	2	Monitored	3,4	Respondents	1.53%	60.5%	86.9%	2.7%	0.0%	10.4%
						Nonrespondents	0.52%	30.3%	100%	0.0%	0.0%	0.0%

---

* Age class 1 consists of students born in September 1981 or earlier. All other students are in age class 2.

Note that invariably within a cell the size of the population represented by the nonrespondents is relatively small. Thus it is not likely in any state that substantial nonresponse bias could be arising from the nonresponse within a single cell. Rather, if such bias is occurring, it must be aggregated across a number of cells having varying characteristics except perhaps for the fact that they involve students of above average age. The small number of nonrespondents within each cell (often as few as five or six) makes it difficult to compare the characteristics of nonrespondents with those of respondents and to characterize the nonrespondents' distributions of gender, race/ethnicity, and median household income.

Of particular note in this table is the fact that all but one of the cells with adjustment factors in excess of 1.25 are from Oklahoma. This occurs because Oklahoma is the only state that required written parental consent before a selected student could participate in the assessment. This requirement resulted in much greater student nonresponse overall than in other states. What the results in Table 7-6 suggest is that this nonresponse is very widely distributed across the various adjustment classes, and is not concentrated among particular types of students. This lessens (but does not eliminate) the likelihood that the relatively high level of student nonresponse in Oklahoma has introduced substantial nonresponse bias. On the other hand, it can be seen, for example, that the percentage of students who are White is consistently several percentage points higher among the nonrespondents than the respondents across classes in Oklahoma. This is reflected in the results in Table 7.3, where the original sample percentage of White students is 72.81, 0.46 percent greater than the weighted sample of respondents (72.35). Before attempting to interpret this slight difference, however, one should note with caution that the reporting of race/ethnicity for assessed students is by the students themselves, whereas for absent students race/ethnicity is reported by school personnel.

7.5 Variation in Weights

After completion of the weighting steps, an analysis was conducted of the distribution of the final student weights in each state. The analysis was intended to check that the various weight components had been derived properly in each state and to examine the impact of the variability of the sample weights on the precision of the sample estimates, both for the state as a whole and for major subgroups within the state.

The analysis was conducted by looking at the distribution of the final student weights, both for the approximately 2,500 assessed students in each state, and for subgroups defined by age, gender, race/ethnicity, level of urbanization, and level of parents' education. Two key aspects of the distribution were considered in each case: the coefficient of variation (equivalently, the relative variance) of the weight distribution; and the presence of outliers (i.e., cases whose weights were several standard deviations away from the median weight).

It was important to examine the coefficient of variation of the weights because a large coefficient of variation reduces the effective size of the sample. Assuming that the variables of interest for individual students are uncorrelated with the weights of the students, the sampling variance of an estimated average or aggregate is approximately $\{1+\left(\frac{C}{100}\right)^2\}$ times as great as the corresponding sampling variance based on a self-weighting sample of the same size, where C is the coefficient of variation of the weights expressed as a percent. Outliers, or cases with extreme weights, were examined because the presence of such an outlier was an indication of the possibility that an error was made in the weighting procedure, and because it was likely that a few extreme cases would contribute substantially to the size of the coefficient of variation.

In most states, the coefficients of variation were 35 percent or less, both for the whole sample and for all major subgroups. This means that the quantity $\{1+\left(\frac{C}{100}\right)^2\}$ was generally below 1.1, and the variation in sampling weights had little impact on the precision of sample estimates.

Large student weights were observed in a few states. These extreme weights generally affected those students in schools for which the grade enrollment available at the time of sample selection proved to be several-fold short of the actual enrollment. An evaluation was made of the impact of trimming these largest weights back to a level consistent with the remaining large weights found in the state. Such a procedure produced some reduction in the size of the coefficient of variation. It was sufficiently modest in each case, however, that we judged that the potential for the introduction of bias through trimming, when combined with the considerable effort required to implement an appropriate trimming procedure, was such that it was preferable not to apply any trimming to the weights in these states. The analyses conducted confirmed that weight components had been calculated and combined correctly, and it was concluded that weight trimming should not be undertaken. Note, however, that weight trimming of school base weights had already been applied in a few cases, prior to the analyses discussed here (see section 7.2.2).

7.6 Calculation of Replicate Weights

A method known as jackknife replication was used to estimate the sampling variance of statistics derived from the full sample. The process of replication involves repeatedly selecting portions of the sample to calculate the statistic of interest; the resultant estimates are known as replicate estimates. The variability among the calculated replicate estimates is then used to obtain the sampling variance of the full-sample estimate. The process of forming the replicate estimates is described below.

7.6.1 Defining Replicate Groups for Variance Estimation

To form replicates for variance estimation, the sampled clusters in each Cluster Type 2 or 3 state (that is, those states where not all schools were selected) were sorted by monitor status, new-school status within monitor status, and finally by selection order within new-school status. The selection order used to form the replicate groups reflected the implicit stratification used in the selection of the sample of schools (see section 3.4.4). Within the sorted file, the basic algorithm for forming the replicate groups was to pair successive clusters, separately within the two monitor status categories. A monitored cluster was always paired with a monitored cluster, and an unmonitored cluster was always paired with an unmonitored cluster. All members (schools) of a cluster received the same pair code, and a substitute school received the pair code of the school it replaced. Double-session substitute schools were in effect assigned two pair codes, one corresponding to the original participating school and the other corresponding to the refusing school for which the extra sessions were conducted.

Since the schools in the Cluster Type 1 states were certainty schools, they were sorted and paired differently. First, each school was assigned a "half-group" code corresponding to the expected number of students selected from the school. For Delaware and the District of Columbia, the value of the half-group code was set to 1 if the expected number of sample students in the school was less than 90; otherwise, the value of the half-group code was set to 2. For schools in Guam, the values of the half-group code ranged from 2 to 8, depending on the estimated grade enrollment of the school; for schools in the Virgin Islands, the values of the half-group code ranged from 2 to 16, depending on the estimated grade enrollment of the school. After assignment of the half-group codes, the schools within each Cluster Type 1 state were sorted by monitor status, half-group code (descending order) within monitor status, and by the estimated grade enrollment of the school within half-group code. Note that the half-group code essentially specifies the number of variance estimation units to be created from the school. For example, two clusters of students (i.e., variance-estimation units) were created from each school having a half-group code of 2, four clusters of students (i.e., variance-estimation units) were created from each school having a half-group code of 4; and so on. Each variance-estimation unit was a systematic sample of students within the school, and successive variance-estimation units in the sorted file were paired to define the replicates.

In some instances, there were an odd number of clusters (in the case of Cluster Type 2 or 3 states) or variance-estimation units (in the case of Cluster Type 1 states) within a monitor-status category. If this occurred, the last "pair" within the monitor-status category actually consisted of three clusters or variance-estimation units. In general, a single replicate was defined by randomly dropping a member (i.e., either a cluster or variance-estimation unit) of a given pair and then reweighting the remaining sample elements to compensate for the dropped unit. If the pair consisted of three units, two groups of two units each were randomly retained to form two replicates.

The number of replicates formed in this manner depended on the number of pairs formed. Based on statistical and computer processing requirements, it was decided that 56 replicates would be sufficient for the variance calculations. In a few states, there were more than 56 initial pairs using the procedures described above. In these states, it was necessary to combine some of the initial replicate groups to reduce the total number of replicates. In general, the goal was to combine an initial pair with another pair consisting of dissimilar schools within the same monitor-status category.

In some states, fewer than 56 replicates were formed. In order to provide a uniform total of 56 replicates, additional sets of replicate weights were created simply by setting the additional sets equal to the set of full-sample weights. This procedure is unbiased and produces appropriate jackknifed sampling errors, while giving uniformity across states in the number of replicate weights.

7.6.2 School-level Replicate Weights

As mentioned above, each replicate sample had to be reweighted to compensate for the dropped unit(s) defining the replicate. For the Cluster Type 2 and 3 states, this reweighting was done in two stages. At the first stage, the *i*th school included in a particular replicate *r* was assigned a replicate-specific school base weight defined as follows:

$$W_{(r)i}^{sch} = K_r W_i^{sch}$$

where W_i^{sch} is the full-sample base weight for school i, and

$$K_r = \begin{cases} 1.5 & \text{if school } i \text{ was contained in a "pair" consisting of 3 units from which the complementary member was dropped to form replicate } r, \\ 2 & \text{if school } i \text{ was contained in a pair consisting of 2 units from which the complementary member was dropped to form replicate } r, \\ 0 & \text{if school } i \text{ was dropped to form replicate } r, \\ 1 & \text{otherwise.} \end{cases}$$

Using the replicate-specific school base weights, $W_{(r)i}^{sch}$, the school-level nonresponse weighting adjustments as described in section 7.3.3 were recalculated for each replicate *r*. That is, the school-level nonresponse adjustment factor for schools in replicate *r* and adjustment class *h* was computed as:

$$F_{(r)h}^{(1)} = \frac{\sum_{i \in C_h} W_{(r)hi}^{sch} \; {}_{hi}}{\sum_{i \in C_h} W_{(r)hi}^{sch} \; {}_{hi} \; {}_{(r)hi}}$$

where

C_h = the subset of school records in adjustment class *h*;

$W_{(r)hi}^{sch}$ = the replicate-*r* base weight of the *i*th school in class *h*;

E_{hi} = the QED grade enrollment for the *i*th school in class *h*;

$$\delta_{(r)hi} = \begin{cases} 1 & \text{if the } i\text{th school in replicate } r \text{ and adjustment class } h \text{ participated in the assessments; and} \\ 0 & \text{otherwise.} \end{cases}$$

The replicate-specific nonresponse-adjusted school weight for the *i*th school in class *h* in replicate *r* was then computed as:

$$W^{adj}_{(r)hi} = F^{(1)}_{(r)h} W^{sch}_{(r)hi} \ .$$

7.6.3 Student-level Replicate Weights

For the Cluster Type 2 and 3 states, replicate-specific adjusted student base weights were calculated by multiplying the replicate-specific adjusted school weights as described above by the corresponding within-school student weights. That is, following the procedures in section 7.3.5, the adjusted student base weight for the *j*th student in adjustment class *k* in replicate *r* was initially computed as:

$$W^{(2)}_{(r)kij} = W^{adj}_{(r)hi} W^{within}_{ij}$$

where

$W^{(2)}_{(r)hi}$ = the nonresponse-adjusted school weight for school *i* in school adjustment class *h* and replicate *r*;

W^{within}_{ij} = the within-school weight for the *j*th student in school *i*.

For the Cluster Type 1 states, the school-level nonresponse adjustment was not replicated since the schools in such states were selected with certainty. In this case, the replicate-specific adjusted student base weight for the *j*th student in adjustment class *k* in replicate *r* was calculated as:

$$W^{(2)}_{(r)kij} = W^{adj}_{hi} W^{within}_{(r)ij}$$

where

W^{adj}_{hi} = the overall nonresponse-adjusted school weight for school/hit *i* in school adjustment class *h*;

$W^{within}_{(r)ij}$ = the replicate-specific within-school weight for the *j*th student in school *i*

= $K_r W^{within}_{ij}$

The factor K_r in the above expression for the replicate-specific within-school weight compensates for the units dropped out in any given replicate (see section 7.6.1) and is defined by:

$$K_r = \begin{cases} 2 & \text{for the students in school } i \text{ who were in a pair from which the complementary variance-estimation unit was dropped to form replicate } r, \\ 0 & \text{for the students in the variance-estimation unit that was dropped to form replicate } r, \\ 1 & \text{otherwise.} \end{cases}$$

The final replicate-specific student weights were then obtained by applying the student nonresponse adjustment procedures (see section 7.3.5) to each set of replicate student weights. Let $F^{(2)}_{(r)k}$ denote the student-level nonresponse adjustment factor for replicate r and adjustment class k. For the Cluster Type 2 and 3 states, the final replicate-r student weight for student j in school i in adjustment class k was calculated as:

$$W^{final}_{(r)kij} = F^{(2)}_{(r)k} W^{adj}_{(r)hi} W^{within}_{ij} \quad .$$

For the Cluster Type 1 states, the corresponding final replicate-r student weight for student j in school i in adjustment class k was calculated as:

$$W^{final}_{(r)kij} = F^{(2)}_{(r)k} W^{adj}_{hi} W^{within}_{ij} \quad .$$

Estimates of the variance of sample-based estimates were calculated as follows:

Let $\hat{x} = \sum_{i=j}^{n} W^{final}_{kij} x_{kij}$ denote an estimated total based on the full sample, and let $\hat{x}_{(r)}$ denote the corresponding estimate based on replicate r. The jackknife variance estimate of $\hat{x}$ was calculated as:

$$var_{JK}(\hat{x}) = \sum_{r=1}^{R} (\hat{x}_{(r)} - \hat{x})^2 \ ,$$

where R is the number of replicates.

7.7 Calculation of School Weights

The school weights described in section 7.3.3 can be used to estimate school-level characteristics and aggregates. However, these school weights are not appropriate as described because small schools had a chance of being selected for either the reading assessment or the mathematics assessment, but not both. To compensate for this factor, school base weights were recomputed appropriately for the full set of schools, regardless of the subject for which they were selected. Using these new base weights, school nonresponse adjustment factors, final school weights, and school replicate weights were derived using the procedures described above.

Chapter 8

THEORETICAL BACKGROUND AND PHILOSOPHY OF NAEP SCALING PROCEDURES

Eugene G. Johnson, Robert J. Mislevy, and Neal Thomas

Educational Testing Service

8.1 OVERVIEW

The primary method by which results from the Trial State Assessment are disseminated is scale-score reporting. With scaling methods, the performance of a sample of students in a subject area or subarea can be summarized on a single scale or series of subscales even when different students have been administered different items. This chapter presents an overview of the scaling methodologies employed in the analyses of the data from NAEP surveys in general and from the Trial State Assessment in reading in particular. Details of the scaling procedures specific to the Trial State Assessment are presented in Chapter 9.

8.2 BACKGROUND

The basic information from an assessment consists of the responses of students to the items presented in the assessment. For NAEP, these items are generated to measure performance on sets of objectives developed by nationally representative panels of learning area specialists, educators, and concerned citizens. Satisfying the objectives of the assessment and ensuring that the tasks selected to measure each goal cover a range of difficulty levels typically requires many items. The Trial State Assessment in reading required 85 items at grade 4. To reduce student burden, each assessed student was presented only a fraction of the full pool of items through multiple matrix sampling procedures.

The most direct manner of presenting the assessment results is to report percent correct statistics for each item. However, because of the vast amount of information, separate results for each of the items in the assessment pool hinders the comparison of the general performance of subgroups of the population. Item-by-item reporting ignores overarching similarities in trends and subgroup comparisons that are common across items.

It is useful to view the assessed items as random representatives of a conceptually infinite pool of items within the same domain and of the same type. In this random item concept, a set of items is taken to represent the domain of interest. An obvious measure of achievement within a domain of interest is the average percent correct across all presented items within that domain. The advantage of averaging is that it tends to cancel out the effects of peculiarities in items that can affect item difficulty in unpredictable ways. Furthermore, averaging makes it possible to compare more easily the general performances of subpopulations.

Despite their advantages, there are a number of significant problems with average percent correct scores. First, the interpretation of these results depends on the selection of the items; the selection of easy or difficult items could make student performance appear to be overly high or low. Second, the average percent correct metric is related to the particular items comprising the average, so that direct comparisons in performance between subpopulations require that those subpopulations have been administered the same set of items. Third, because this approach limits comparisons to percents correct on specific sets of items, it provides no simple way to report trends over time when the item pool changes. Finally, direct estimates of statistics such as the proportion of students who would respond correctly to 80 percent of the items in the pool are not possible when every student is administered only a fraction of the item pool. While the mean percent correct across all items in the pool can be readily obtained (as the average of the individual item percent correct statistics), distributional statistics, such as quantiles of the distribution of scores across the full set of items, cannot be readily obtained without additional assumptions.

These limitations can be overcome by the use of response scaling methods. If several items require similar skills, the regularities observed in response patterns can often be exploited to characterize both respondents and items in terms of a relatively small number of variables. These variables include a respondent-specific variable, called proficiency, which quantifies a respondent tendency to answer items correctly and item-specific variables which indicate characteristics of the item such as its difficulty, ability to distinguish between individuals with different levels of proficiency, and the chances of a very low proficiency respondent correctly answering the item. (These variables are discussed in more detail in the next section). When combined through appropriate mathematical formulas, these variables capture the dominant features of the data. Furthermore, all students can be placed on a common scale, even though none of the respondents take all of the items within the pool. Using the scale, it becomes possible to discuss distributions of proficiency in a population or subpopulation and to estimate the relationships between proficiency and background variables.

It is important to point out that any procedure of aggregation, from a simple average to a complex multidimensional scaling model, highlights certain patterns at the expense of other potentially interesting patterns that may reside within the data. Every item in a NAEP survey is of interest and can provide useful information about what young Americans know and can do. The choice of an aggregation procedure must be driven by a conception of just which patterns are salient for a particular purpose.

The scaling for the Trial State Assessment in reading was carried out separately within the two reading content areas specified in the framework for grade 4 reading. This scaling within subareas was done because it was anticipated that different patterns of performance might exist for these essential subdivisions of the subject area. The two content area scales correspond with two purposes of reading—Reading for Literary Experience and Reading to Gain Information. By creating a separate scale for each of these content areas, potential differences in [illegible]ation performance between the content areas are maintained.

The creation of a series of separate scales to describe reading performance does not preclude the reporting of an overall reading composite as a single index of overall reading performance. A composite is computed as the weighted average of the two content area scales, where the weights correspond to the relative importance given to each content area as defined by the framework. The composite provides a global measure of performance within the subject area, while the constituent content area scales allow the measurement of important interactions within educationally relevant subdivisions of the subject area.

8.3 SCALING METHODOLOGY

This section reviews the scaling models employed in the analyses of data from the Trial State Assessment in reading and the 1992 national reading assessment, and the multiple imputation or "plausible values" methodology that allows such models to be used with NAEP's sparse item-sampling design. The reader is referred to Mislevy (1991) for an introduction to plausible values methods and a comparison with standard psychometric analyses, to Mislevy, Johnson and Muraki (1992) and Beaton and Johnson (1992) for additional information on how the models are used in NAEP, and to Rubin (1987) for the theoretical underpinnings of the approach. It should be noted that the imputation procedure used by NAEP is a mechanism for providing plausible values for proficiencies and not for filling in blank responses to background variables.

While the NAEP procedures were developed explicitly to handle the characteristics of NAEP data, they build on other research, and are paralleled by other researchers. See, for example Dempster, Laird, and Rubin (1977); Little and Rubin (1983, 1987); Andersen (1980); Engelen (1987); Hoijtink (1991); Laird (1978); Lindsey, Clogg, and Grego (1991); Zwinderman (1991); Tanner and Wong (1987); and Rubin (1991).

The 85 reading items administered at grade 4 in the Trial State Assessment were also administered to fourth-grade students in the national reading assessment. However, because the administration procedures differed, the Trial State Assessment data were scaled independently from the national data. The national data also included results for students in grades 8 and 12. Details of the scaling of the Trial State Assessment and the subsequent linking to the results from the national reading assessment are provided in Chapter 9.

8.3.1 The Scaling Models

Three distinct scaling models were used in the analysis of the data from the Trial State Assessment. Each of the models are based on item response theory (IRT; e.g., Lord, 1980). Each is a "latent variable" model, defined separately for each of the scales, and quantifying respondents' tendencies to provide correct answers to the items contributing to a scale as a function of a parameter that is not directly observed, called proficiency on the scale.

A three-parameter logistic (3PL) model was used for the multiple-choice items. The fundamental equation of the 3PL model is the probability that a person whose proficiency on scale *k* is characterized by the *unobservable* variable θ_k will respond correctly to item *j*:

$$\begin{aligned} P(X_j = 1|\theta_k, a_j, b_j, c_j) &= C_j + \frac{(1 - C_j)}{1 + \exp[-1.7a_j\,(\theta_k - b_j)]} \\ &\equiv P_{j1}(\theta_k) \ , \end{aligned} \tag{8.1}$$

where

x_j	is the response to item *j*, 1 if correct and 0 if not;
a_j	where $a_j > 0$, is the slope parameter of item *j*, characterizing its sensitivity to proficiency;
b_j	is the threshold parameter of item *j*, characterizing its difficulty; and
c_j	where $0 \leq c_j < 1$, is the lower asymptote parameter of item *j*, reflecting the chances of students of very low proficiency selecting the correct option.

Further define the probability of an incorrect response to the item as

$$P_{j0} = P(x_j = 0|\theta_k, a_j, b_j, c_j) = 1 - P_{j1}(\theta_k) \tag{8.2}$$

A two-parameter logistic (2PL) model was used for short constructed-response items, which were scored correct or incorrect. The form of the 2PL model is the same as equations (8.1) and (8.2) with the c_j parameter fixed at zero.

In addition to the multiple-choice and short constructed-response items, eight extended constructed-response items were presented in the Trial State and grade 4 national assessments. Each of these items was scored on a multipoint scale with potential scores ranging from 0 to 4. Items that are scored on a multipoint scale are referred to as polytomous items, in contrast with the multiple-choice and short constructed-response items, which are scored correct/incorrect and referred to as dichotomous items.

The polytomous items were scaled using a generalized partial credit model (Muraki, 1992). The fundamental equation of this model is the probability that a person with proficiency θ_k on scale *k* will have, for the *j*th polytomous item, a response x_j that is scored in the *i*th of m_j ordered score categories:

$$P(X_j = i \mid \theta_k, a_j, b_j, d_{j,1}, \ldots, d_{j,m_j-1}) = \frac{\exp\left(\sum_{v=0}^{i} 1.7a_j(\theta_k - b_j + d_{j,v})\right)}{\sum_{g=0}^{m_j-1} \exp\left(\sum_{v=0}^{g} 1.7a_j(\theta_k - b_j + d_{j,v})\right)} \tag{8.3}$$

$$\equiv P_{ji}(\theta_k)$$

where

m_j	is the number of categories in the response to item *j*
x_j	is the response to item *j*, with possibilities $0,1,\ldots,m_j-1$
a_j	is the slope parameter;
b_j	is the item location parameter characterizing overall difficulty; and
$d_{j,i}$	is the category *i* threshold parameter (see below).

Indeterminacies in the parameters of the above model are resolved by setting $d_{j,0} = 0$ and setting $\sum_{i=1}^{m_j-1} d_{j,i} = 0$. Muraki (1992) points out that $b_j - d_{j,i}$ is the point on the θ_k scale at which the plots of $P_{j,i-1}(\theta_k)$ and $P_{ji}(\theta_k)$ intersect and so characterizes the point on the θ_k scale at which the response to item *j* has the highest probability of incurring a change from response category *i*-1 to *i*.

When $m_j = 2$, so that there are two score categories (0,1), it can be shown that $P_{ji}(\theta_k)$ of equation 8.3 for $i = 0,1$ corresponds respectively to $P_{i0}(\theta_k)$ and $P_{j1}(\theta_k)$ of the 2PL model (equations 8.1 and 8.2 with $c_j = 0$).

A typical assumption of item response theory is the conditional independence of the response by an individual to a set of items, given the individual's proficiency. That is, conditional on the individual's θ_k, the joint probability of a particular response pattern $\underset{\sim}{x} = (x_1, \ldots, x_n)$ across a set of *n* items is simply the product of terms based on (8.1), (8.2), and (8.3):

$$P(\underset{\sim}{x} \mid \theta_k, \textit{item parameters}) = \prod_{j=1}^{n} \prod_{i=0}^{m_j-1} P_{ji}(\theta_k)^{u_{ji}} \tag{8.4}$$

where $P_{ji}(\theta_k)$ is of the form appropriate to the type of item (dichotomous or polytomous), m_j is taken equal to 2 for the dichotomously scored items, and u_{ji} is an indicator variable defined by

$$u_{ji} = \begin{cases} 1 \text{ if response } x_j \text{ was in category } i \\ 0 \text{ otherwise.} \end{cases}$$

It is also typically assumed that response probabilities are conditionally independent of background variables ($\underset{\sim}{y}$), given θ_k, or

$$P(\underset{\sim}{x}|\theta_k, item\ parameters, \underset{\sim}{y}) = p(\underset{\sim}{x}|\theta_k, item\ parameters) \tag{8.5}$$

After $\underset{\sim}{x}$ has been observed, equation 8.4 can be viewed as a likelihood function, and provides a basis for inference about θ_k or about item parameters. Estimates of item parameters were obtained by the NAEP BILOG/PARSCALE program, which combines Mislevy and Bock's (1982) BILOG and Muraki and Bock's (1991) PARSCALE computer programs, and which concurrently estimates parameters for all items (dichotomous and polytomous). The item parameters are then treated as known in subsequent calculations. The parameters of the items constituting each of the separate scales were estimated independently of the parameters of the other scales. Once items have been calibrated in this manner, a likelihood function for the scale proficiency θ_k is induced by a vector of responses to any subset of calibrated items, thus allowing θ_k-based inferences from matrix samples.

In all NAEP IRT analyses, missing responses at the end of each block a student was administered were considered "not-reached," and treated as if they had not been presented to the respondent. Missing responses to dichotomous items before the last observed response in a block were considered intentional omissions, and treated as fractionally correct at the value of the reciprocal of the number of response alternatives. These conventions are discussed by Mislevy and Wu (1988). With regard to the handling of not-reached items, Mislevy and Wu found that ignoring not-reached items introduces slight biases into item parameter estimation to the degree that not-reached items are present and speed is correlated with ability. With regard to omissions, they found that the method described above provides consistent limited-information likelihood estimates of item and ability parameters under the assumption that respondents omit only if they can do no better than responding randomly.

Although the IRT models are employed in NAEP only to summarize performance, a number of checks are made to detect serious violations of the assumptions underlying the models (such as conditional independence). When warranted, remedial efforts are made to mitigate the effects of such violations on inferences. These checks include comparisons of empirical and theoretical item response functions to identify items for which the IRT model may provide a poor fit to the data.

Scaling areas in NAEP are determined *a priori* by considerations of content as collections of items for which overall performance is deemed to be of interest, as defined by the frameworks developed by the National Assessment Governing Board. A proficiency scale θ_k is defined *a priori* by the collection of items representing that scale. What is important, therefore, is that the models capture salient variation in the response data to effectively summarize the overall performance on the content area of the populations and subpopulations being assessed. Because of the *a priori* definition of the latent proficiency variable, departure from conditional independence tends to cancel out over items and does not seriously affect the estimation of whole group and subpopulation distributions, except when substantial differential item functioning (DIF) is found simultaneously for many items. NAEP has routinely conducted DIF analyses to guard against potential biases in making subpopulation comparisons based on the proficiency distributions.

The local independence assumption embodied in equation 8.4 implies that item response probabilities depend only on θ and the specified item parameters, and not on the position of the item in the booklet, the content of items around an item of interest, or the test-administration timing conditions. However, these effects are certainly present in any application. The practical question is whether inferences based on the IRT probabilities obtained via 8.4 are robust with respect to the ideal assumptions underlying the IRT model. Our experience with the 1986 NAEP reading anomaly (Beaton & Zwick, 1990) has shown that for measuring small changes over time, changes in item context and speededness conditions can lead to unacceptably large random error components. These can be avoided by presenting items used to measure change in identical test forms, with identical timings and administration conditions. Thus, we do *not* maintain that the item parameter estimates obtained in any particular booklet configuration are appropriate for other conceivable configurations. Rather, we assume that the parameter estimates are context-bound. (For this reason, we prefer common population equating to common item equating whenever equivalent random samples are available for linking.) This is the reason that the data from the Trial State Assessment were calibrated separately from the data from the national NAEP—since the administration procedures differed somewhat between the Trial State Assessment and the national NAEP, the values of the item parameters could be different. Chapter 9 provides details on the procedures used to link the results of the 1992 Trial State Assessment to those of the 1992 national assessment.

8.3.2 An Overview of Plausible Values Methodology

Item response theory was developed in the context of measuring individual examinees' abilities. In that setting, each individual is administered enough items (often 60 or more) to permit precise estimation of his or her θ, as a maximum likelihood estimate $\hat{\theta}$, for example. Because the uncertainty associated with each θ is negligible, the distribution of θ, or the joint distribution of θ with other variables, can then be approximated using individuals' $\hat{\theta}$ values as if they were θ values.

This approach breaks down in the assessment setting when, in order to provide broader content coverage in limited testing time, each respondent is administered relatively few items in a scaling area. The problem is that the uncertainty associated with individual θs is too large to ignore, and the features of the $\hat{\theta}$ distribution can be seriously biased as estimates of the θ distribution. (The failure of this approach was verified in early analyses of the 1984 NAEP reading survey; see Wingersky, Kaplan, & Beaton, 1987.) "Plausible values" were developed as a way to estimate key population features consistently, and approximate others no worse than standard IRT procedures would. A detailed development of plausible values methodology is given in Mislevy (1991). Along with theoretical justifications, that paper presents comparisons with standard procedures, discussions of biases that arise in some secondary analyses, and numerical examples.

The following provides a brief overview of the plausible values approach, focusing on its implementation in the Trial State Assessment analyses.

Let $\underline{y}$ represent the responses of all sampled examinees to background and attitude questions, along with design variables such as school membership, and let $\underline{\theta}$ represent the subscale proficiency values. If $\underline{\theta}$ were known for all sampled examinees, it would be possible to compute a statistic $t(\underline{\theta},\underline{y})$—such as a subscale or composite subpopulation sample mean, a sample percentile point, or a sample regression coefficient—to estimate a corresponding population quantity T. A function $U(\underline{\theta},\underline{y})$—e.g., a jackknife estimate—would be used to gauge sampling uncertainty, as the variance of t around T in repeated samples from the population.

Because the scaling models are latent variable models, however, $\underline{\theta}$ values are not observed even for sampled students. To overcome this problem, we follow Rubin (1987) by considering $\underline{\theta}$ as "missing data" and approximate $t(\underline{\theta},\underline{y})$ by its expectation given $(\underline{x},\underline{y})$, the data that actually were observed, as follows:

$$\begin{aligned} t^*(\underline{x},\underline{y}) &= E[t(\underline{\theta},\underline{y})\,|\,\underline{x},\underline{y}] \\ &= \int t(\underline{\theta},\underline{y})\, p(\underline{\theta}\,|\,\underline{x},\underline{y})\, d\underline{\theta}\ . \end{aligned} \tag{8.6}$$

It is possible to approximate t^* using random draws from the conditional distribution of the scale proficiencies given the item responses x_i, background variables y_i, and model parameters for sampled student i. These values are referred to as "imputations" in the sampling literature, and "plausible values" in NAEP. The value of $\underline{\theta}$ for any respondent that would enter into the computation of t is thus replaced by a randomly selected value from their conditional distribution. Rubin (1987) proposes that this process be carried out several times—"multiple imputations"—so that the uncertainty associated with imputation can be quantified. The average of the results of, for example, M estimates of t, each computed from a different set of plausible values, is a Monte Carlo approximation of (8.6); the variance among them, B, reflects uncertainty due to not observing θ, and must be added to the estimated expectation of $U(\underline{\theta},\underline{y})$, which reflects uncertainty due to testing only a sample of students from the population. Section 8.5 explains how plausible values are used in subsequent analyses.

It cannot be emphasized too strongly that **plausible values are *not* test scores for *individuals*** in the usual sense. Plausible values are offered only as intermediary computations for calculating integrals of the form of equation 8.6, in order to estimate *population* characteristics. When the underlying model is correctly specified, plausible values will provide consistent estimates of population characteristics, even though they are not generally unbiased estimates of the proficiencies of the individuals with whom they are associated. The key idea lies in a contrast between plausible values and the more familiar θ estimates of educational measurement that are in some sense optimal for each examinee (e.g., maximum likelihood estimates, which are consistent estimates of an examinee's θ, and Bayes estimates, which provide minimum mean-squared errors with respect to a reference population): *Point estimates that are optimal for individual examinees have distributions that can produce decidedly nonoptimal (specifically, inconsistent) estimates of population characteristics* (Little & Rubin, 1983). Plausible values, on the other hand, are constructed explicitly to provide consistent estimates of population effects.

8.3.3 Computing Plausible Values in IRT-based Scales

Plausible values for each respondent i are drawn from the conditional distribution $p(\underline{\theta}_i|x_i,y_i,\Gamma,\Sigma)$, where Γ and Σ are regression model parameters defined in this subsection. This subsection describes how, in IRT-based scales, these conditional distributions are characterized, and how the draws are taken. An application of Bayes' theorem with the IRT assumption of conditional independence produces

$$p(\underline{\theta}_i|x_i,y_i,\Gamma,\Sigma) \propto P(x_i|\underline{\theta}_i,y_i,\Gamma,\Sigma)\, p(\underline{\theta}_i|y_i,\Gamma,\Sigma) = P(x_i|\underline{\theta}_i)\, p(\underline{\theta}_i|y_i,\Gamma,\Sigma) \ , \tag{8.7}$$

where, for vector-valued $\underline{\theta}_i$, $P(x_i|\underline{\theta}_i)$ is the product over scales of the *independent likelihoods* induced by responses to items within each scale, and $p(\underline{\theta}_i|y_i,\Gamma,\Sigma)$ is the multivariate—and generally nonindependent—*joint density* of proficiencies for the scales, conditional on the observed value y_i of background responses, and the parameters Γ and Σ. The scales are determined by the item parameter estimates that constrain the population mean to zero and standard deviation to one. The item parameter estimates are fixed and regarded as population values in the computation described in this subsection.

In the analyses of the data from the Trial State Assessment and the data from the national reading assessment, a normal (Gaussian) form was assumed for $p(\underline{\theta}_i|y_i\Gamma,\Sigma)$, with a common variance, Σ, and with a mean given by a linear model with slope parameters, Γ, based on the first 99 to 162 principal components of 361 selected main-effects and two-way interactions of the complete vector of background variables. The included principal components will be referred to as the *conditioning variables*, and will be denoted y^c. (The complete set of original background variables used in the Trial State Assessment reading analyses are listed in Appendix C.) The following model was fit to the data within each state:

$$\theta = \Gamma' y^c + \varepsilon \ , \tag{8.8}$$

where ε is normally distributed with mean zero and variance Σ. The number of principal components of the conditioning variables used for each state was sufficient to account for 90 percent of the total variance of the full set of conditioning variables (after standardizing each variable). As in regression analysis, Γ is a matrix each of whose columns is the *effects* for one scale and Σ is the matrix *variance of residuals* between subscales. By fitting the model (8.8) separately within each state, interactions between each state and the conditioning variables are automatically included in the conditional joint density of scale proficiencies.

Maximum likelihood estimates of Γ and Σ, denoted by $\hat{\Gamma}$ and $\hat{\Sigma}$, are obtained from Sheehan's (1985) MGROUP computer program using the EM algorithm described in Mislevy (1985). The EM algorithm requires the computation of the mean, $\bar{\underline{\theta}}_i$, and variance, Σ_i^p, of the posterior distribution in (8.7). These moments are computed using higher order asymptotic corrections (Thomas, 1992).

After completion of the EM algorithm, the plausible values are drawn in a three-step process from the joint distribution of the values of Γ for all sampled respondents. First, a value of Γ is drawn from a normal approximation to $P(\Gamma,\Sigma|x_i,y_i)$ that fixes Σ at the value $\hat{\Sigma}$, (Thomas,

BEST COPY AVAILABLE

1992). Second, conditional on the generated value of Γ (and the fixed value of $\Sigma = \hat{\Sigma}$), the mean, $\bar{\underline{\theta}}_i$, and variance, Σ_i^p, of the posterior distribution in equation 8.7 (i.e., $p(\underline{\theta}_i | x_i, y_i, \Gamma, \Sigma)$) are computed using the same methods applied in the EM algorithm. In the third step, the $\underline{\theta}_i$ are drawn independently from a multivariate normal distribution with mean $\bar{\underline{\theta}}_i$ and variance Σ_i^p, approximating the distribution in (8.7). These three steps are repeated five times producing five imputations of $\bar{\underline{\theta}}_i$ for each sampled respondent.

8.4 ACHIEVEMENT LEVELS

Since its beginning, a goal of NAEP has been to inform the public about what students in American schools know and can do. While the NAEP scales provide information about the distributions of proficiency for the various subpopulations, they do not directly provide information about the meaning of various points on the scale. Traditionally, meaning has been attached to educational scales by norm-referencing—that is, by comparing students at a particular scale level to other students. In contrast, NAEP achievement levels describe selected points on the scale in terms of the types of skills that are or should be exhibited by students scoring at that level. The achievement level process was applied to the 1992 national NAEP reading composite. However, since the Trial State Assessment scales were linked to the national scales, the interpretations of the selected levels also apply to the Trial State Assessment.

The National Assessment Governing Board has determined that achievement levels shall be the first and primary way of reporting NAEP results. Setting achievement levels is a method for setting standards on the NAEP assessment that identify what students should know and be able to do at various points on the reading composite. For each grade, three levels were defined—basic, proficient, and advanced. Based on initial policy definitions of these levels, panelists were asked to determine operational descriptions of the levels appropriate with the content and skills assessed in the reading assessment. With these descriptions in mind, the panelists were then asked to rate the assessment items in terms of the expected performance of marginally acceptable examinees at each of these three levels. These ratings were then mapped onto the NAEP scale to obtain the achievement level cutpoints for reporting. Further details of the achievement level-setting process appear in Appendix F.

The achievement level-setting process specifies expected performance of students at each of the three achievement levels. To determine the types of skills currently exhibited by students at each of the levels, ETS applied a modified anchoring procedure to the 1992 reading achievement levels. As applied to the achievement levels, the anchoring process was designed to determine the sets of questions that students scoring at or above each achievement level cutpoint could perform with a high degree of success. Specifically, a question was identified as anchoring at an achievement level for a given grade if it was answered correctly by at least 65 percent of the students in that grade scoring at the cutpoint of that achievement level, and by less than 65 percent of the students scoring at the cutpoints for any lower achievement level. A committee of reading experts, educators, and others was assembled to review the questions and, using their knowledge of reading and student performance, to generalize from the questions to descriptions of the types of skills exhibited at each achievement level. Further details of the anchoring process appear in Appendix G.

8.5 ANALYSES

When survey variables are observed without error from every respondent, standard variance estimators quantify the uncertainty associated with sample statistics from the only source of the uncertainty, namely the sampling of respondents. Item percents correct for NAEP cognitive items meet this requirement, but scale-score proficiency values do not. The IRT models used in their construction posit an unobservable proficiency variable θ to summarize performance on the items in the subarea. The fact that θ values are not observed even for the respondents in the sample requires additional statistical analyses to draw inferences about θ distributions and to quantify the uncertainty associated with those inferences. As described above, Rubin's (1987) multiple imputations procedures were adapted to the context of latent variable models to produce the plausible values upon which many analyses of the data from the Trial State Assessment were based. This section describes how plausible values were employed in subsequent analyses to yield inferences about population and subpopulation distributions of proficiencies.

8.5.1 Computational Procedures

Even though one does not observe the θ value of respondent *i*, one does observe variables that are related to it: x_i, the respondent's answers to the cognitive items he or she was administered in the area of interest, and y_i, the respondent's answers to demographic and background variables. Suppose one wishes to draw inferences about a number $T(\underline{\theta},\underline{Y})$ that could be calculated explicitly if the θ and y values of each member of the population were known. Suppose further that if θ values were observable, we would be able to estimate T from a sample of N pairs of θ and y values by the statistic $t(\underline{\theta},\underline{y})$ [where $(\underline{\theta},\underline{y}) \equiv (\theta_1,y_1,\ldots,\theta_N,y_N)$], and that we could estimate the variance in t around T due to sampling respondents by the function $U(\underline{\theta},\underline{y})$. Given that observations consist of (x_i,y_i) rather than $(\underline{\theta}_i,y_i)$, we can approximate t by its expected value conditional on $(\underline{x},\underline{y})$, or

$$t^*(\underline{x},\underline{y}) = E[t(\underline{\theta},\underline{y})|\underline{x},\underline{y}] = \int t(\underline{\theta},\underline{y})\, p(\underline{\theta}|\underline{x},\underline{y})\, d\underline{\theta} .$$

It is possible to approximate t^* with random draws from the conditional distributions $p(\underline{\theta}_i|x_i,y_i)$, which are obtained for all respondents by the method described in section 8.3.3. Let $\underline{\theta}_m$ be the *m*th such vector of "plausible values," consisting of a multidimensional value for the latent variable of each respondent. This vector is a plausible representation of what the true $\underline{\theta}$ vector might have been, had we been able to observe it.

The following steps describe how an estimate of a scalar statistic $t(\underline{\theta},\underline{y})$ and its sampling variance can be obtained from M (>1) such sets of plausible values. (Five sets of plausible values are used in NAEP analyses of the Trial State Assessment.)

1) Using each set of plausible values $\underset{\sim}{\theta}_m$ in turn, evaluate t as if the plausible values were true values of $\underset{\sim}{\theta}$. Denote the results $\hat{t}_m$, for $m = 1,...,M$.

2) Using the jackknife variance estimator defined in Chapter 7, compute the estimated sampling variance of $\hat{t}_m$, denoting the result U_m.

3) The final estimate of t is

$$t^* = \sum_{m=1}^{M} \frac{\hat{t}_m}{M}.$$

4) Compute the average sampling variance over the M sets of plausible values, to approximate uncertainty due to sampling respondents:

$$U^* = \sum_{m=1}^{M} \frac{U_m}{M}.$$

5) Compute the variance among the M estimates $\hat{t}_m$, to approximate uncertainty due to not observing θ values from respondents:

$$B_M = \sum_{m=1}^{M} \frac{(\hat{t}_m - t^*)^2}{(M - 1)}$$

6) The final estimate of the variance of t^* is the sum of two components:

$$V = U^* + (1 + M^{-1}) B_M.$$

Note: Due to the excessive computation that would be required, NAEP analyses did not compute and average jackknife variances over all five sets of plausible values, but only on the first set. Thus, in NAEP reports, U^* is approximated by U_1.

8.5.2 Statistical Tests

Suppose that if θ values were observed for sampled students, the statistic $(t - T)/U^{1/2}$ would follow a t-distribution with d degrees of freedom. Then the incomplete-data statistic $(t^* - T)/V^{1/2}$ is approximately t-distributed, with degrees of freedom given by

$$\nu = \frac{1}{\dfrac{f_M^2}{M - 1} + \dfrac{(1 - f_M)^2}{d}}$$

where f_M is the proportion of total variance due to not observing θ values:

$$f_M = (1 + M^{-1}) B_M / V_M.$$

When B_M is small relative to U^*, the reference distribution for incomplete-data statistics differs little from the reference distribution for the corresponding complete-data statistics. This is the case with main NAEP reporting variables. If, in addition, d is large, the normal approximation can be used to flag "significant" results.

For k-dimensional t, such as the k coefficients in a multiple regression analysis, each U_m and U^* is a covariance matrix, and B_M is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity $(T-t^*)\ V^{-1}\ (T-t^*)'$ is approximately F distributed, with degrees of freedom equal to k and ν, with ν defined as above but with a matrix generalization of f_M:

$$f_M = (1+M^{-1})\ Trace\ (B_M V_M^{-1})/k\ .$$

By the same reasoning as used for the normal approximation for scalar t, a chi-square distribution on k degrees of freedom often suffices.

8.5.3 Biases in Secondary Analyses

Statistics t^* that involve proficiencies in a scaled content area and variables included in the conditioning variables y^c are consistent estimates of the corresponding population values T. Statistics involving background variables y that were *not* conditioned on, or relationships among proficiencies from *different* content areas, are subject to asymptotic biases whose magnitudes depend on the type of statistic and the strength of the relationships of the nonconditioned background variables to the variables that were conditioned on and to the proficiency of interest. That is, the large sample expectations of certain sample statistics need not equal the true population parameters.

The *direction* of the bias is typically to underestimate the effect of nonconditioned variables. For details and derivations see Beaton and Johnson (1990), Mislevy (1991), and Mislevy and Sheehan (1987, section 10.3.5). For a given statistic t^* involving one content area and one or more nonconditioned background variables, the *magnitude* of the bias is related to the extent to which observed responses x account for the latent variable θ, and the degree to which the nonconditioned background variables are explained by conditioning background variables. The first factor—conceptually related to test reliability—acts consistently in that greater measurement precision reduces biases in *all* secondary analyses. The second factor acts to reduce biases in certain analyses but increase it in others. In particular,

- High shared variance between conditioned and nonconditioned background variables *mitigates* biases in analyses that involve only proficiency and nonconditioned variables, such as marginal means or regressions.

- High shared variance *exacerbates* biases in regression coefficients of conditional effects for nonconditioned variables, when nonconditioned and conditioned background variables are analyzed jointly as in multiple regression.

The large number of background variables that have been included in the conditioning vector for the Trial State Assessment allows a large number of secondary analyses to be carried out with little or no bias, and mitigates biases in analyses of the marginal distributions of θ in nonconditioned variables. Kaplan and Nelson's analysis of the 1988 NAEP reading data (some results of which are summarized in Mislevy, 1991), which had a similar design and fewer conditioning variables, indicate that the potential bias for nonconditioned variables in multiple regression analyses is below 10 percent, and biases in simple regression of such variables is below 5 percent. Additional research (summarized in Mislevy, 1990) indicates that most of the bias reduction obtainable from conditioning on a large number of variables can be captured by instead conditioning on the first several principal components of the matrix of all original conditioning variables. This procedure was adopted for the Trial State Assessment by replacing the conditioning effects by the first *K* principal components, where *K* was selected so that 90 percent of the total variance of the full set of conditioning variables (after standardization) was captured. Mislevy (1990) shows that this puts an upper bound of 10 percent on the average bias for all analyses involving the original conditioning variables.

Chapter 9

DATA ANALYSIS AND SCALING FOR THE 1992 TRIAL STATE ASSESSMENT IN READING[1]

Nancy L. Allen, John Mazzeo, Steven P. Isham, Y. Fai Fong, and Drew W. Bowker

Educational Testing Service

9.1 OVERVIEW

This chapter describes the analyses carried out in the development of the 1992 Trial State Assessment reading scales. The procedures used were similar to those employed in the analysis of the 1990 and 1992 Trial State Assessments in mathematics (Mazzeo, 1991 and Mazzeo, Chang, Kulick, Fong, & Grima, 1993) and are based on the philosophical and theoretical underpinnings described in the previous chapter.

There were five major steps in the analysis of the Trial State Assessment reading data, each of which is described in a separate section:

- conventional item and test analyses (section 9.3);
- item response theory (IRT) scaling (section 9.4);
- estimation of state and subgroup proficiency distributions based on the "plausible values" methodology (section 9.5);
- linking of the 1992 Trial State Assessment scales to the corresponding scales from the 1992 national assessment (section 9.6); and
- creation of the Trial State Assessment reading composite scale (section 9.7).

To set the context within which to describe the methods and results of scaling procedures, a brief review of the assessment instruments and administration procedures is provided.

---

[1] Thanks to James Carlson, Huahua Chang, John Donoghue, David Freund, Angela Grima, Laura Jerry, Eugene Johnson, Edward Kulick, Jennifer Nelson, and Spencer Swinton for their help in completing the analysis. Thanks also to Huahua Chang and Angela Grima for their contributions to the original draft of this chapter.

9.2 DESCRIPTION OF ITEMS, ASSESSMENT BOOKLETS, AND ADMINISTRATION PROCEDURES

The 1992 Trial State Assessment in reading was administered to fourth-grade public-school students only. The items in the instruments were based on the curriculum framework described in Chapter 2. The instruments included many more constructed-response items than were previously included in NAEP reading assessments. Some of these items were scored dichotomously due to the short length of the responses expected. One item per block was an extended constructed-response item. Each extended constructed-response item required about five minutes to complete and was scored by specially trained readers on a 0-to-4 scale. (The scoring process is described in detail in Chapter 5.) During the scaling process, the categories in the 0-to-4 scale were collapsed for one of these items, and the 0 (off-task) category was treated as "not administered" for each of the items so that the scaling model used for these items fit the data more closely. The extended constructed-response items appeared in varying positions within each block. These items, including the recoding of the 0-to-4 scale, are described in more detail in section 9.4.1.

The fourth-grade item pool contained 85 items. They were categorized into one of two content areas: 43 items for Reading for Literary Experience and 42 for Reading to Gain Information. These items, consisting of 42 multiple-choice items, 35 short constructed-response items, and 8 extended constructed-response items, were divided into 8 mutually exclusive blocks. The composition of each block of items, in terms of content and format, is given in Table 9-1. Note that each block contained items from only one of the two content domains.

The 8 blocks were used to form 16 different booklets according to a partially balanced incomplete block (PBIB) design (see Chapter 2 for details). Each of these booklets contained two blocks of items, and each block of items appeared in exactly four booklets. To balance possible block position effect, each block appeared twice as the first block of reading items and twice as the second block. In addition, the design required that each block of items be paired in a booklet with every other block of items in the same content domain exactly once. Finally, each block of items was included in a booklet with a block of items from the other area.

Within each administration site, all booklets were "spiraled" together in a random sequence and distributed to students sequentially, in the order of the students' names on the Student Listing Form (see Chapter 4). As a result of the partial BIB design and the spiraling of booklets, a considerable degree of balance was achieved in the data collection process. Each block of items (and, therefore, each item) was administered to randomly equivalent samples of students of approximately equal size (i.e., about 4/16 or 1/4 of the total sample size) within each jurisdiction and across all jurisdictions. In addition, within and across jurisdictions, randomly equivalent samples of approximately equal size received each particular block of items as the first or second block within a booklet.

As described in Chapter 4, a randomly selected half of the administration sessions within each state were observed by Westat-trained quality control monitors. Thus, within and across states, randomly equivalent samples of students received each block of items in a particular position within a booklet under monitored and unmonitored administration conditions.

Table 9-1

1992 NAEP Reading Block Composition by Scale and Item Type for Grade 4*

	Reading for Literary Experience				Reading to Gain Information				Total			
Block	**Multiple Choice**	**Short Constructed Response**	**Extended Constructed Response**	**Total**	**Multiple Choice**	**Short Constructed Response**	**Extended Constructed Response**	**Total**	**Multiple Choice**	**Short Constructed Response**	**Extended Constructed Response**	**Total**
R3	6	4	1	11	0	0	0	0	6	4	1	11
R4	5	6	1**	12	0	0	0	0	5	6	1	12
R5	7	3	1	11	0	0	0	0	7	3	1	11
R6	0	0	0	0	5	4	1	10	5	4	1	10
R7	0	0	0	0	4	5	1	10	4	5	1	10
R8	0	0	0	0	5	4	1	10	5	4	1	10
R9	4	4	1	9	0	0	0	0	4	4	1	9
R10	0	0	0	0	6	5	1***	12	6	5	1	12
Total	22	17	4	43	20	18	4	42	42	35	8	85

---

* At grade 4, each block contained one reading passage.

** Two categories of response for this item were collapsed during the scaling process.

*** This item appears in the final position in the block.

9.3 ITEM ANALYSES

9.3.1 Conventional Item and Test Analyses

Table 9-2 contains summary statistics for each block of items. Block-level statistics are provided both overall and by serial position of the block within booklet. To produce these tables, data from all 44 jurisdictions were aggregated and statistics were calculated using rescaled versions of the final sampling weights provided by Westat. The rescaling, carried out within each jurisdiction, constrained the sum of the sampling weights within that jurisdiction to be equal to its sample size. The sample sizes for each jurisdiction were approximately equal[2]. Use of the rescaled weights does nothing to alter the value of statistics calculated separately within each jurisdiction. However, for statistics obtained from samples that combine students from different jurisdictions, use of the rescaled weights results in a roughly equal contribution of each jurisdiction's data to the final value of the estimate. As discussed in Mazzeo (1991), equal contribution of each jurisdiction's data to the results of the IRT scaling was viewed as a desirable outcome and, as described in the scaling section below, these same rescaled weights were only adjusted slightly in carrying out that scaling. Hence, the item analysis statistics shown in Table 9-2 are approximately consistent with the weighting used in scaling. The original final sampling weights provided by Westat were used in reporting.

Table 9-2 shows the number of students assigned each block of items, the average item score, the average polyserial correlation, and the proportion of students attempting the last item in the block. The average item score for the block is the average, over items, of the score means for each of the individual items in the block. For binary-scored multiple-choice and constructed-response items, these score means correspond to the proportion of students who correctly answered each item. For the extended constructed-response items, the score means were calculated as item score mean divided by the maximum number of points possible.

In NAEP analyses (both conventional and IRT-based), a distinction is made between missing responses at the end of each block (i.e., missing responses subsequent to the last item the student answered) and missing responses prior to the last observed response. Missing responses before the last observed response are considered intentional omissions. In calculating the average score for each item, only students classified as having been presented the item were included in the denominator of the statistic. Intentional omissions were treated as incorrect responses. Missing responses at the end of the block are considered "not-reached," and treated as if they had not been presented to the student. The proportion of students attempting the last item of a block (or, equivalently, 1 minus the proportion of students not reaching the last item) is often used as an index of the degree of speededness associated with the administration of that block of items.

Standard practice at ETS is to treat all nonrespondents to the last item as if they had not reached the item. For multiple-choice and standard constructed-response items, the use of such a convention most often produces a reasonable pattern of results in that the proportion reaching the last item is not dramatically smaller than the proportion reaching the next-to-last item.

---

[2] The size of the sample in Guam was approximately 1/3 the sample size for the other jurisdictions.

Table 9-2

Descriptive Statistics for Each Block of Items by Position Within Test Booklet and Overall

Statistic	Position	R3	R4	R5	R6	R7	R8	R9	R10
Unweighted sample size	1	13635	13661	13597	13823	13937	13972	13485	13866
	2	13850	13718	13468	13863	13838	13948	13424	13664
	All	27485	27379	27065	27686	27775	27920	26909	27530
Average item score	1	.63	.68	.46	.56	.43	.51	.49	.63
	2	.63	.66	.44	.54	.42	.49	.47	.61
	All	.63	.67	.45	.55	.42	.50	.48	.62
Average r-polyserial	1	.71	.68	.61	.60	.67	.61	.68	.62
	2	.75	.71	.63	.62	.70	.62	.72	.65
	All	.73	.70	.62	.61	.68	.62	.70	.64
Proportion of students attempting last item	1	.67	.56	.68	.61	.56	.68	.65	.76
	2	.84	.74	.83	.81	.71	.81	.82	.87
	All	.75	.65	.75	.71	.63	.75	.73	.82

However, for the blocks that ended with extended constructed-response items, use of the standard ETS convention resulted in an extremely large drop in the proportion of students attempting the final item. A drop of such magnitude seemed somewhat implausible. Therefore, for blocks ending with an extended constructed-response items, students who answered the next-to-last item but did not respond to the extended constructed-response item were classified as having intentionally omitted the last item.

The average polyserial correlation is the average, over items, of the item-level polyserial correlations (r-polyserial). For each item-level r-polyserial, the total block number-correct score (including the item in question, and with students receiving zero points for all not-reached items) was used as the criterion variable for the correlation. For dichotomously scored items, the item-level r-polyserial correlations are standard r-biserial correlations. Data from students classified as not reaching the item were omitted from the calculation of the statistic.

As is evident from Table 9-2, the difficulty and the internal consistency of the blocks varied somewhat. Such variability was expected since these blocks were not created to be parallel in either difficulty or content. Based on the proportion of students attempting the last item, all of the blocks seem to be somewhat speeded. Only 65 percent of the students receiving block R4 and 63 percent of the students receiving block R7 reached the last item in the block.

This table also indicates that there was little variability in average item scores or average polyserial correlations for each block by serial position within the assessment booklet. The differences in item statistics were small for items in appearing in blocks in the first position and in the second position. However, differences were consistent in their direction. Average item scores were highest when each block was presented in the first position. Average polyserial correlations were highest when each block was presented in the second position. An aspect of block-level performance that did differ noticeably by serial position was the proportion of students attempting the last item in the block. As shown in Table 9-2, the percentage of the students attempting the last item increased as the serial position of the block increased. Students may have learned to pace themselves through the later block after they had experienced the format of the first block they received. A study was completed to examine the effect of the serial position differences on scaling. Due to the partial BIB design of the booklets, those effects were minimal.

As mentioned earlier, in an attempt to maintain rigorous standardized administration procedures across the states, a randomly selected 50 percent of all sessions within each state was observed by a Westat-trained quality control monitor. Observations from this random half of the sessions provided information about the quality of administration procedures and the frequency of departures from standardized procedures in the monitored sessions (see Chapter 4, section 4.3.6 for a discussion of the results of these observations). In addition, unexpectedly large differences in results from monitored and unmonitored sessions (i.e., differences larger than those to be expected due to sampling fluctuation) provided the means to identify instances of cheating, breaches of test security, or other breaks in standardization occurring in the unmonitored sessions that might threaten the validity of assessment results.

When results were aggregated over all participating jurisdictions, there was little difference between the performance of students who attended monitored or unmonitored sessions. The average item score (over all 8 blocks and over all 44 participating jurisdictions) was .56 for both monitored and unmonitored sessions. Table 9-3 provides, for each block of items, the average item score, average r-polyserial, and the proportion of students attempting the last item for students whose sessions were monitored and students whose sessions were not monitored. Little or no differences by session type were evident. These aggregate results are quite consistent with those observed in the 1990 and 1992 Trial State Assessment in mathematics, where no evidence was found that students who attended monitored sessions performed differently than those who attended unmonitored sessions.

Figure 9-1 presents stem-and-leaf displays of the differences between monitored and unmonitored average item scores (over all 8 blocks) on each of the two content area scales for each of the 44 jurisdictions participating in the 1992 Trial State Assessment. Stem-and-leaf displays, developed by Tukey (1977), are somewhat like histograms. For this figure (and other stem-and-leaf displays that follow), the first column contains observation depths (Hoaglin, Mosteller, & Tukey, 1983). Depths are essentially cumulative frequencies, counted up from the lowest value for score intervals ("stems") below the median and counted down from the highest value for score intervals above the median. The second column contains a count of the number of "leaves" on each stem. In histogram terms, these counts would be considered frequencies. The remainder of the figure contains the stem-and-leaf display. The combination of a stem with each of its leaves gives the actual value of one observation (i.e., the difference in average item scores for monitored and unmonitored sessions in a participating jurisdiction).

The median differences (monitored minus unmonitored) were .004 and .006 respectively for the Reading for Literary Experience scale and the Reading to Gain Information scale. In evaluating the magnitude of these differences, it should be noted that the standard error for a difference in proportions from independent simple random samples of size 1,250 (half the typical total state sample size of 2,500) from a population with a true proportion of .5 is about .02. For samples with complex sampling designs like NAEP, the standard errors tend to be larger than those associated with simple random sampling. A reasonable estimate of the design effect for proportion correct statistics based on past NAEP experience is about 1.5 (Johnson & Rust, 1992), which suggests that a typical estimate of the standard error of the difference between monitored and unmonitored sessions would be about .024. On the Reading for Literary Experience scale the absolute differences in item score means for 37 of the 44 participating states were less than .02, and all but five were less than .024. The differences with the largest magnitude were positive, with values of .031 and .038. On the Reading to Gain Information scale, the absolute differences in item score means for 35 of the 44 participants were less than .02. The differences with the largest magnitudes were .030, .032, and .040. In summary, differences in results obtained from the two types of sessions at the fourth grade were within the bounds expected due to sampling fluctuation.

9.3.2 Differential Item Functioning (DIF) Analyses

Prior to scaling, differential item functioning (DIF) analyses were carried out on 1992 NAEP reading data from the national cross-sectional samples at grades 4, 8, and 12 and the Trial State Assessment sample at grade 4. The purpose of these analyses was to identify items that were differentially difficult for various subgroups and to reexamine such items with respect to their fairness and their appropriateness for inclusion in the scaling process. The information in this section focuses mainly on the analyses conducted on the Trial State Assessment data. A

Table 9-3

Block-level Descriptive Statistics for Monitored and Unmonitored Sessions

Statistic	R3	R4	R5	R6	R7	R8	R9	R10
Unweighted sample size								
Unmonitored	13774	13654	13552	13839	13833	13934	13423	13732
Monitored	13711	13725	13513	13847	13942	13986	13486	13798
Average item score								
Unmonitored	.63	.67	.45	.55	.42	.50	.48	.62
Monitored	.63	.67	.45	.55	.43	.50	.48	.62
Average r-polyserial								
Unmonitored	.73	.70	.61	.61	.69	.62	.71	.64
Monitored	.72	.69	.63	.60	.68	.62	.69	.63
Proportion of students attempting last item								
Unmonitored	.75	.65	.75	.71	.64	.74	.73	.82
Monitored	.75	.65	.76	.71	.63	.75	.73	.81

Figure 9-1

Stem-and leaf Display* of State-by-state Differences in Average Item Scores by Scale (Monitored Minus Unmonitored)

READING FOR LITERARY EXPERIENCE

N = 44, Median = 0.004, Quartiles = -0.0055, 0.0125
Decimal point is 2 places to the left of the colon

```
 3    3  -2 : 520
10    7  -1 : 9874200
17    7  -0 : 7432222
     13   0 : 0013446777888
14   10   1 : 0014445778
 4    2   2 : 45
 2    2   3 : 18
```

READING TO GAIN INFORMATION

N = 44, Median = 0.006, Quartiles = -0.003, 0.0135
Decimal point is 2 places to the left of the colon

```
 2    2  -2 : 32
 7    5  -1 : 94331
13    6  -0 : 764333
     15   0 : 001233556677799
16    9   1 : 001134688
 7    4   2 : 0345
 3    2   3 : 02
 1    1   4 : 0
```

---

* The first column of numbers shows observation depths; the second column shows the number of observations; the remainder of the figure contains the stem-and-leaf display.

description of the results based on the national assessment appears in the technical report for that assessment.

The DIF analyses were based on the Mantel-Haenszel chi-square procedure, as adapted by Holland and Thayer (1988). The procedure tests the statistical hypothesis that the odds of correctly answering an item are the same for two groups of examinees that have been matched on some measure of proficiency (usually referred to as the matching criterion). The groups being compared are often referred to as the focal group (usually a minority or other group of interest, such as Black examinees or female examinees) and the reference group (usually White examinees or male examinees). The measure of proficiency used is typically the number-correct score on some collection of items. Separate analyses were performed for each block of items (i.e., data were pooled across booklets containing the block being analyzed), and number-correct score on the block of items in question was used as the measure of proficiency.

For each item in the assessment, an estimate was produced of the Mantel-Haenszel common odds-ratio, expressed on the ETS delta scale for item difficulty . The estimates indicate the difference between reference group and focal group item difficulties (measured in ETS delta scale units), and typically run between about +3 and -3. Positive values indicate items that are differentially easier for the focal group than the reference group after making an adjustment for the overall level of proficiency in the two groups. Similarly, negative values indicate items that are differentially harder for the focal group than the reference group. It is common practice at ETS to categorize each item into one of three categories (Petersen, 1988): "A" (items exhibiting no DIF), "B" (items exhibiting a weak indication of DIF), or "C" (items exhibiting a strong indication of DIF). Items in category A have Mantel-Haenszel values that do not differ significantly from 0 at the alpha = .05 level. Two conditions must be met in order for items to fall in category B. The Mantel-Haenszel value for the item must: (1) be significantly greater than 0 but not significantly greater than 1 at the .05 level and (2) must be less than 1.5 in absolute magnitude. Category C items are those with Mantel-Haenszel values that are significantly greater than 1 and larger than 1.5 in absolute magnitude.

For each block of items at grade 4 a single set of analyses was carried out based on equal-sized random samples of data from all participating jurisdictions. Each set of analyses involved four reference group/focal group comparisons: male/female, White/Asian American, White/Black, and White/Hispanic.

All analyses used rescaled sampling weights. A separate rescaled weight was defined for each comparison as:

$$\textit{Rescaled Weight} = \textit{Original Weight} \times \frac{\textit{Total Sample Size}}{\textit{Sum of the Weights}}$$

where the total sample size is the total number of students for the two groups being analyzed (e.g., for the White/Hispanic comparison, the total number of White and Hispanic examinees in the sample at that grade), and the sum of the weights is the sum of the sampling weights of all the students in the sample for the two groups being analyzed. Four rescaled weights were computed for White examinees—one for the gender comparison and three for the race/ethnicity comparisons. Two rescaled overall weights were computed for the Asian

American, Black, and Hispanic examinees—one for the gender comparison and another for the appropriate race/ethnicity comparison.

The ETS generalized program IANA83 was used to carry out the DIF analyses. Two-sided modification[3] was used. In the calculation of number-correct scores for the matching criterion, both not-reached and omitted items were considered as wrong responses. For each item, calculation of the Mantel-Haenszel statistic did not include data from examinees who did not reach the item in question. Because the Mantel-Haenszel procedure, as currently implemented, is appropriate only for dichotomously scored items, the extended constructed-response items had to be scored dichotomously for the DIF analyses. Extended constructed responses rated as "essential" or "extensive" were scored as correct; all other responses were scored as incorrect.

At grade 4, 85 items were analyzed. Table 9-4 provides a summary of the results of the DIF analyses for the collection of items grouped by content area. The table provides two sets of five frequency distributions for the categorized Mantel-Haenszel statistics for the items in each of the scales. The leftmost frequency distribution gives the number (and percent) of items in each of five categories (C+, B+, A, B-, C-) based on the largest absolute DIF value obtained for the item across the four reference group/focal group comparisons that were carried out. The remaining four frequency distributions give the number of items with indices in each DIF category for each of the four reference group/focal group comparisons.

Two items were classified as "C" items for at least one of the analyses for the fourth-grade Trial State Assessment data. One of the "C" items was differentially more difficult for the Asian American examinees than for the White examinees. This item was an item measuring Reading for Literary Experience. The other "C" item was in the Reading to Gain Information scale and was differentially more difficult for the White examinees than for the Asian American examinees. There were also more items categorized as "B" items in White/Asian American comparisons than in any of the other comparisons.

Following standard practice at ETS for DIF analyses conducted on final test forms, all "C" items were reviewed by a committee of trained test developers and subject-matter specialists. Such committees are charged with making judgments about whether or not the differential difficulty of an item is *unfairly* related to group membership. As pointed out by Zieky (1993):

> It is important to realize that *DIF* is not a synonym for *bias*. The item response theory based methods, as well as the Mantel-Haenszel and standardization methods of DIF detection, will identify questions that are not measuring the same dimension(s) as the bulk of the items in the matching criterion....Therefore, judgement is required to determine whether or not the difference in difficulty shown by a DIF index is *unfairly* related to group membership. The judgement of fairness is based on whether or not the difference in difficulty is believed to be related to the

---

[3] Modification refers to the procedure in which items classified as "C" items in an initial DIF analysis are deleted from the matching criterion, and a second DIF analysis is run. Two-sided means that "C" items are deleted from the criterion regardless of which group they favor.

Table 9-4

Frequency Distributions of DIF Statistics for Grade 4 Items Grouped by Content Area

Category of Maximum Absolute DIF Value For All Comparisons			Number of Items in Category of DIF Value for Each Comparison (Reference Group/Focal Group)			
DIF Category*	Number	Percent	Male/Female	White/Black	White/Hispanic	White/Asian Amer.
Reading for Literary Experience						
C+	0	0.0	0	0	0	0
B+	8	18.6	1	0	2	6
A	27	62.8	41	41	40	31
B-	7	16.3	1	2	1	5
C-	1	2.3	0	0	0	1
Reading to Gain Information						
C+	1	2.4	0	0	0	1
B+	10	23.8	4	2	1	5
A	22	52.4	37	38	38	31
B-	9	21.4	1	2	3	5
C-	0	0.0	0	0	0	0

---

* Categories are A, B, and C. (+) indicates items in the category that are differentially easier for the focal group; (-) indicates items in the category that are differentially more difficult for the focal group.

> construct being measured....The fairness of an item depends directly on the purpose for which a test is being used. For example, a science item that is differentially difficult for women may be judged to be fair in a test designed for certification of science teachers because the item measures a topic that every entry-level science teacher should know. However, that same item, with the same DIF value, may be judged to be unfair in a test of general knowledge designed for all entry-level teachers. (p. 340)

The committee assembled to review NAEP items included both ETS staff and outside members with expertise in the field. It was the committee's judgment that none of the "C" items for the national or Trial State Assessment data were functioning differentially due to factors irrelevant to test objectives. Hence, none of the items were removed from scaling due to differential item functioning.

9.4 ITEM RESPONSE THEORY (IRT) SCALING

Separate IRT-based scales were developed using the scaling models described in Chapter 8. Two scales were produced by separately calibrating the sets of items classified in each of the two content areas.

Figure 9-2 contains stem-and-leaf displays of the average scores for the items comprising each of the fourth-grade scales. The averages are based on the entire sample of students in the Trial State Assessment and use the same rescaled sampling weights described in the previous section. As a whole, the fourth-grade students found the set of items in the Reading to Gain Information scale to be the most difficult.

For the reasons discussed in Mazzeo (1991), for each scale, a single set of item parameters for each item was estimated and used for all jurisdictions. Item parameter estimation was carried out using a 25 percent systematic random sample of the students participating in the 1992 Trial State Assessment and included equal numbers of students from each participating jurisdiction, half from monitored sessions and half from unmonitored sessions. The sample consisted of 27,632 students, with 628 students being sampled from each of the 44 participating jurisdictions. Of the 628 records sampled from each jurisdiction, 314 were drawn from the monitored sessions and 314 were drawn from the unmonitored sessions. The rescaled weights for the 25 percent sample of students used in item calibration were adjusted slightly to ensure that 1) each states' data contributed equally to the estimation process, and 2) data from monitored and unmonitored sessions contributed equally. For each jurisdiction, the sum of the rescaled sampling weights for the set of monitored and unmonitored records selected for the sample was obtained (these sums are denoted as WM_s and WU_s, respectively). Then, for each jurisdiction, the rescaled weights for individuals in the sample (denoted as W_{si}) were adjusted so that the sum of the weights for the monitored and unmonitored sessions would each be equal to 314. Thus for the monitored students in the sample,

$$W_{si}^{*} = W_{si}(314/WM_s),$$

Figure 9-2

Stem-and-leaf Display* of Average Item Scores by Scale

READING FOR LITERARY EXPERIENCE

N = 43, Median = 0.6, Quartiles = 0.418, 0.738
Decimal point is 1 place to the left of the colon

```
 2    2   1 : 57
 6    4   2 : 0127
 9    3   3 : 036
16    7   4 : 0234468
19    3   5 : 568
     10   6 : 0001133479
14   10   7 : 2234445588
 4    4   8 : 0448
```

READING TO GAIN INFORMATION

N = 42, Median = 0.532, Quartiles = 0.372, 0.694
Decimal point is 1 place to the left of the colon

```
 1    1   0 : 9
 2    1   1 : 7
 4    2   2 : 69
13    9   3 : 114555789
18    5   4 : 22689
      7   5 : 2233368
17    7   6 : 1244469
10    9   7 : 001256689
 1    0   8 :
 1    1   9 : 2
```

---

* The first column of numbers shows observation depths; the second column shows the number of observations; the remainder of the figure contains the stem-and-leaf display.

and for the unmonitored students

$$W_{si}^{*} = W_{si}(314/WU_s),$$

where W_{si}^{*} denotes the adjusted rescaled weight for individual *i* from state *s*. These adjusted rescaled weights for the 25 percent sample of students were used only in item calibration.

As mentioned above, the sample used for item calibration was also constrained to contain an equal number of students from the monitored and unmonitored sessions from each of the participating jurisdictions. To the extent that items may have functioned differently in monitored and unmonitored sessions, the single set of item parameter obtained define a sort of average item characteristic curve for the two types of sessions. Table 9-3 (shown earlier) presented block-level item statistics that suggested little, if any, differences in item functioning by session type. Figure 9-3 presents the results of supplementary analyses organized by scale.

Figure 9-3 contains plots of differences in score means (monitored minus unmonitored) against the score means for the monitored sessions for the items in each of the two scales. The differences between session type appear small on both scales, with a slight tendency for performance to be higher in the monitored sessions.

9.4.1 Item Parameter Estimation

For each content area scale, item parameter estimates were obtained by the NAEP BILOG/PARSCALE program, which combines Mislevy and Bock's (1982) BILOG and Muraki and Bock's (1991) PARSCALE computer programs. The program uses marginal estimation procedures to estimate the parameters of the one-, two-, and three-parameter logistic models, and the generalized partial credit model described by Muraki (1992).

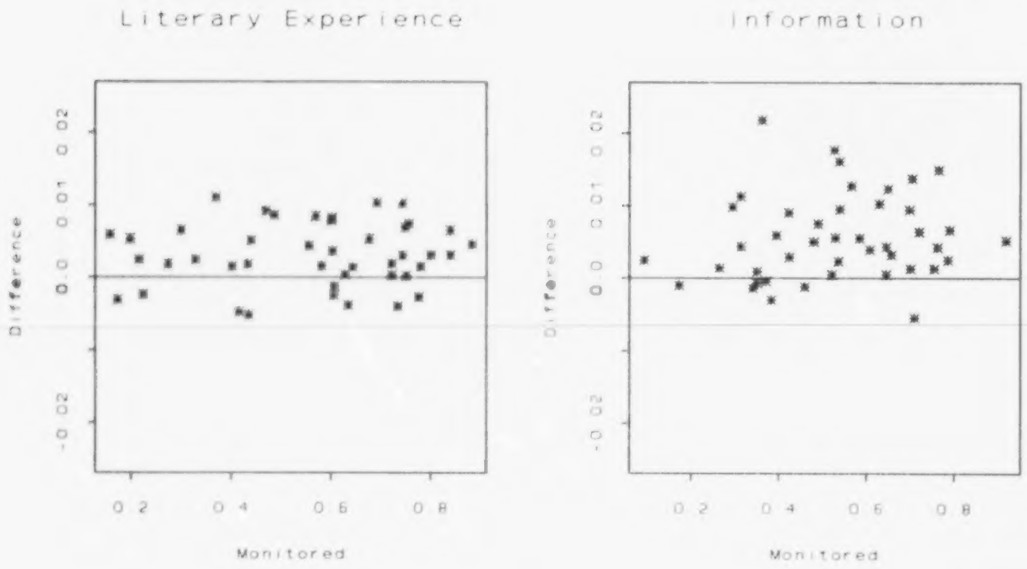
All multiple-choice items were dichotomously scored and were scaled using the three-parameter logistic model. Omitted responses to multiple-choice items were treated as fractionally correct, with the fraction being set to 1 over the number of response options. All short constructed-response items were dichotomously scored and were scaled using the two-parameter logistic model. Omitted responses to short constructed-response items were treated as incorrect.

There were a total of eight extended constructed-response items. Each of these items was also scaled using the generalized partial credit model. Four scoring levels were defined:

0 Unsatisfactory response or omitted;
1 Partial response;
2 Essential response; and
3 Extensive response.

Figure 9-3

Differences in Item Scores (Monitored Minus Unmonitored) Plotted Against Monitored Item Scores



Literary Experience

Information

Not reached and off-task responses were treated as if the item was not administered to the student. Table 9-5 provides a listing of the blocks, positions within the block, content area classifications, and NAEP identification numbers for all extended constructed-response items included in the 1992 assessment.

Bayes modal-estimates of all item parameters were obtained from the BILOG/PARSCALE program. Prior distributions were imposed on item parameters with the following starting values: thresholds (normal [0,2]); slopes (log-normal [0,.5]); and asymptotes (two-parameter beta with parameter values determined as functions of the number of response options for an item and a weight factor of 50). The locations (but not the dispersions) were updated at each program estimation cycle in accordance with provisional estimates of the item parameters.

As was done for the 1990 and 1992 Trial State Assessments in mathematics, item parameter estimation proceeded in two phases. First, the subject ability distribution was assumed fixed (normal [0,1]) and a stable solution was obtained. The parameter estimates from this solution were then used as starting values for a subsequent set of runs in which the subject ability distribution was freed and estimated concurrently with item parameter estimates. After each estimation cycle, the subject ability distribution was re-standardized to have a mean of zero and standard deviation of one. Correspondingly, parameter estimates for that cycle were also linearly re-standardized.

During and subsequent to item parameter estimation, evaluations of the fit of the IRT models were carried out for each of the items in the grade 4 item pools. These evaluations were conducted to determine the final composition of the item pool making up the scales by identifying misfitting items that could not be included. Evaluations of model fit were based primarily on a graphical analysis. For binary-scored items, model fit was evaluated by examining plots of nonmodel-based estimates of the expected conditional (on θ) proportion correct versus the proportion correct predicted by the estimated item characteristic curve (see Mislevy & Sheehan, 1987, p. 302). For the extended constructed-response items, similar plots were produced for each item category characteristic curve.

As with most procedures that involve evaluating plots of data versus model predictions, a certain degree of subjectivity is involved in determining the degree of fit necessary to justify use of the model. There are a number of reasons why evaluation of model fit relied primarily on analyses of plots rather than seemingly more objective procedures based on goodness-of-fit indices such as the "pseudo chi-squares" produced in BILOG (Mislevy & Bock, 1982). First, the exact sampling distributions of these indices when the model fits are not well understood, even for fairly long tests. Mislevy and Stocking (1987) point out that the usefulness of these indices appears particularly limited in situations like NAEP where examinees have been administered relatively short tests. Work reported by Stone, Ankenmann, Lane, and Liu (1993) using simulated data suggests that the correct reference chi-square distributions for these indices have considerably fewer degrees of freedom than the value indicated by the BILOG/PARSCALE program and require additional adjustments of scale. However, it is not yet clear how to estimate the correct number of degrees of freedom and necessary scale factor adjustment factors. Consequently, pseudo chi-square goodness-of-fit indices are used only as rough guides in interpreting the severity of model departures.

Table 9-5

Extended Constructed-response Items
1992 Trial State Assessment in Reading

Block	Position In Block	Scale	NAEP ID
R3	6	Literary Experience	R012006
R4	11	Literary Experience	R012111
R5	7	Literary Experience	R012607
R6	4	Gain Information	R012204
R7	8	Gain Information	R012708
R8	5	Gain Information	R012305
R9	1	Literary Experience	R012401
R10	12	Gain Information	R012512

Second, as discussed in Chapter 8, it is almost certainly the case that, for most items, item-response models hold only to a certain degree of approximation. Given the large samples sizes used in NAEP and the Trial State Assessment, there will be sets of items for which one is almost certain to reject the hypothesis that the model fits the data even though departures are minimal in nature or involve kinds of misfit unlikely to impact on important model-based inferences. In practice, one is almost always forced to temper statistical decisions with judgments about the severity of model misfit and the potential impact of such misfit on final results.

In making decisions about excluding items from the final scales, a balance was sought between being too stringent, hence deleting too many items and possibly damaging the content representativeness of the pool of scaled items, and too lenient, hence including items with model fit poor enough to invalidate the types of model-based inferences made from NAEP results. Items that clearly did not fit the model were not included in the final scales; however, a certain degree of misfit was tolerated for a number of items included in the final scales.

For the large majority of the grade 4 items, the fit of the model was extremely good. Figure 9-4 provides a typical example of what the plots look like for this class of items. The plots that are shown are for items from the Reading for Literary Experience scale. The item at the top of the plot is a multiple-choice item; the item at the bottom of the plot is a binary-scored constructed-response item. In each plot, the y-axis indicates the probability of a correct response and the x-axis indicates proficiency level (theta). The diamonds show estimates of the conditional (on theta) probability of a correct response that do not assume a logistic form (referred to subsequently as nonlogistic-based estimates). The sizes of the diamonds are proportional to the estimated density of the theta distribution at the indicated value. The solid curve shows the estimated item response function. The item response function provides estimates of the conditional probability of a correct response based on an assumed logistic form. The vertical dashed line indicates the estimated location parameter (b) for the item and the horizontal dashed line (top plot only) indicates the estimated lower asymptote (c). Also shown in the plot are the actual values of the item parameter estimates (lower right-hand corner) as well as the proportion of students that answered the item correctly (upper left-hand corner). As is evident from the plots, the nonlogistic-based estimates of conditional probabilities are in extremely close agreement with those given by the estimated item response function.

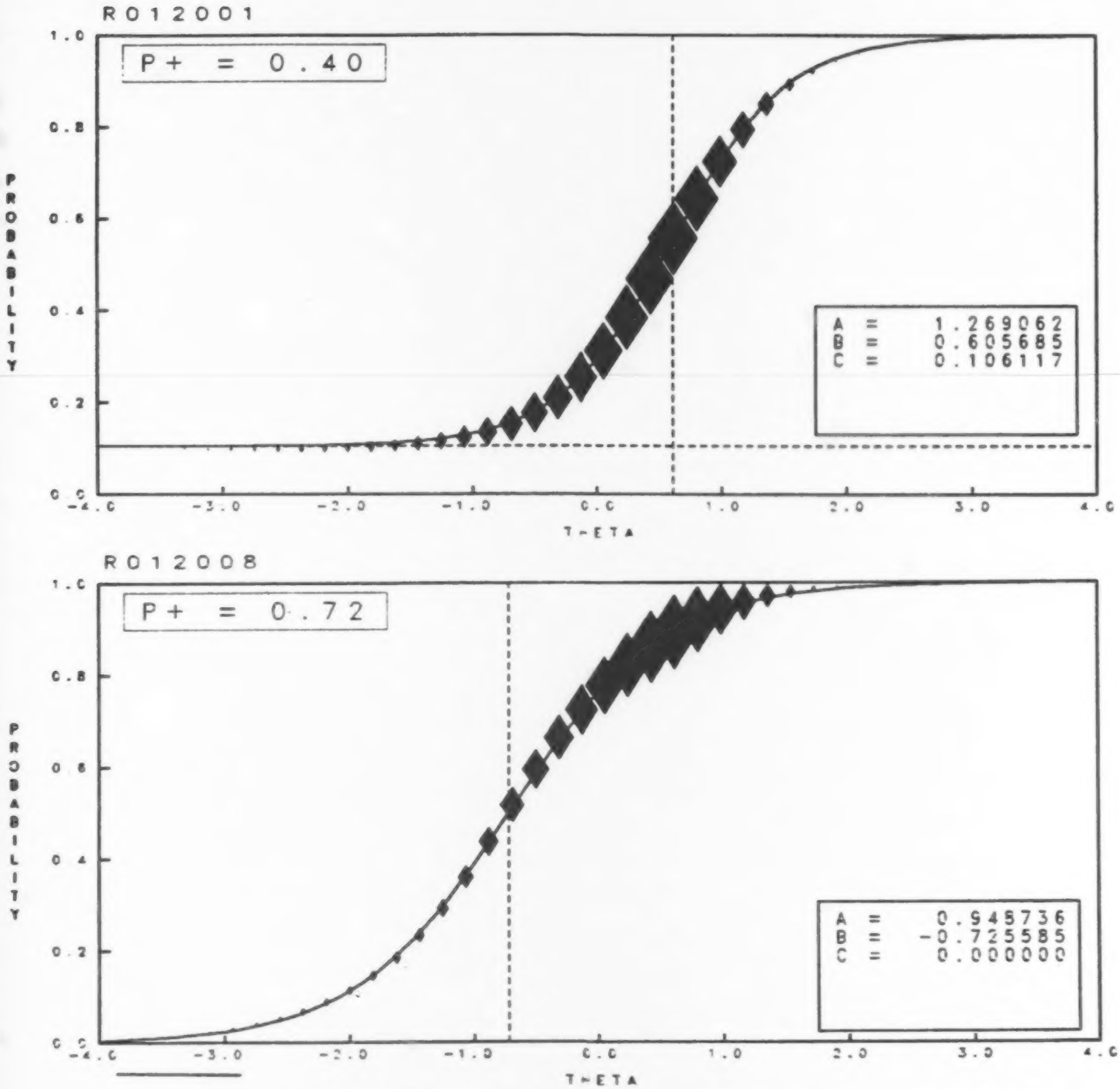
Figure 9-5 provides an example of a plot for a four-category extended constructed-response item exhibiting good model fit. Like the plots for the binary items, this plot shows two estimates of each item category characteristic curve, one set that does not assume the partial credit model (shown as diamonds) and one that does (the solid curves). The dashed vertical lines show the location of the estimated category thresholds for the item (d_1 to d_3; see Chapter 8, sections 8.3.1). The estimates for all parameters for the item in question are also indicated on the plot. As with Figure 9-4, the two sets of estimates agree quite well, although there are slight differences between the two. An aspect of Figure 9-5 worth noting is the large proportion of examinees that responded in the two lowest response categories for this item[4]. Although few

---

[4] This is evidenced by the relatively large size of the diamonds indicating nonlogistic-based estimated conditional probabilities for these two categories.

Figure 9-4

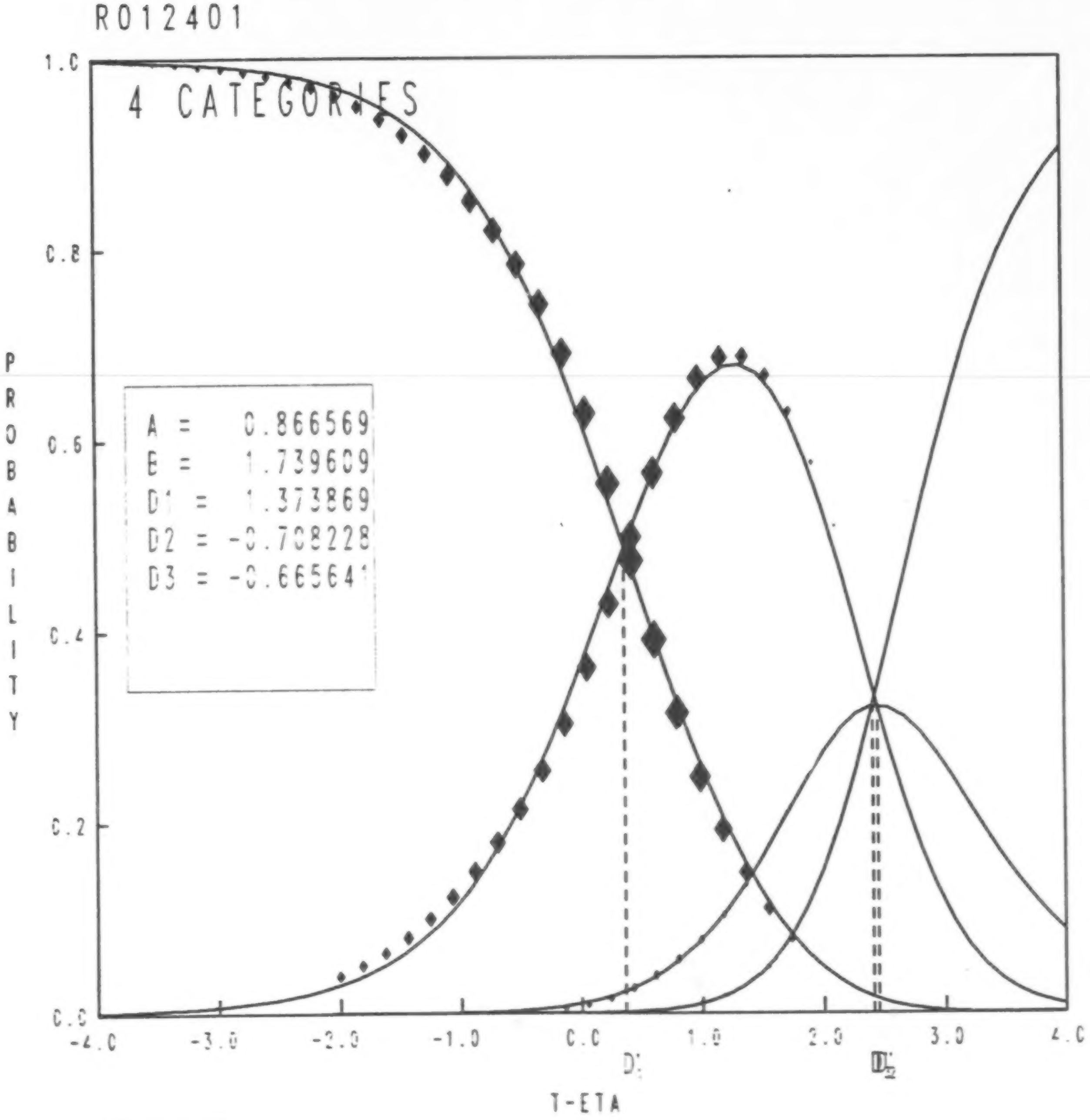
Plots* Comparing Empirical and Model-based Estimates of Item Response Functions for Binary-scored Items Exhibiting Good Model Fit



* Diamonds indicate estimated conditional probabilities obtained without assuming a logistic form; solid curve indicates estimated item response function assuming a logistic form.

Figure 9-5

Plot* Comparing Empirical and Model-based Estimates of Item Category Characteristic Curves for a Polytomously Scored Item Exhibiting Good Model Fit



---

* Diamonds indicate estimated conditional probabilities obtained without assuming a logistic form; solid curve indicates estimated item response function assuming a logistic form.

student responses were categorized in the highest two categories, there were adequate data to estimate the model-based estimates for those categories (the solid curves). Such results were typical for the extended constructed-response items. Substantial proportions of examinees were either unable or unwilling to provide even minimally adequate answers to such items.

As discussed above, some of the items retained for the final scales display some degree of model misfit. Figures 9-6 (binary-scored items) and 9-7 (extended constructed-response item) provide typical examples of such items. In general, good agreement between nonlogistic and logistic estimates of conditional probabilities were found for the regions of the theta scale that includes most of the examinees. Misfit was confined to conditional probabilities associated with theta values in the tails of the subject ability distributions.

Only one item in the assessment received special treatment in the scaling process. The generalized partial credit model did not fit the responses to the extended constructed-response item R012111 well. For this Reading for Literary Experience item, which appeared in the eleventh position in block R4, the categories 0 and 1 were combined and the other categories were relabeled. Therefore the codings for the three scoring levels were defined:

0 Unsatisfactory, partial response, or omitted;
1 Essential response; and
2 Extensive response.

Plots for this item are given in Figures 9-8 and 9-9 before and after collapsing the unsatisfactory and partial response categories. The large differences between the estimates of the category characteristic curves when the partial credit model is assumed (shown as solid curves) and when the model is not assumed (shown as diamonds) indicate that the two lowest categories lack good model fit in Figure 9-8. In contrast, except for the tendency for the nonlogistic-based estimates to be somewhat different from the model-based estimates for theta values greater than 1, Figure 9-9 shows good model fit. Note that this item is functioning primarily as a dichotomous item due to the small frequencies in the top category. There were enough data, however, to calculate the model-based estimates of the category characteristic curve for this category (shown as the rightmost solid curve).

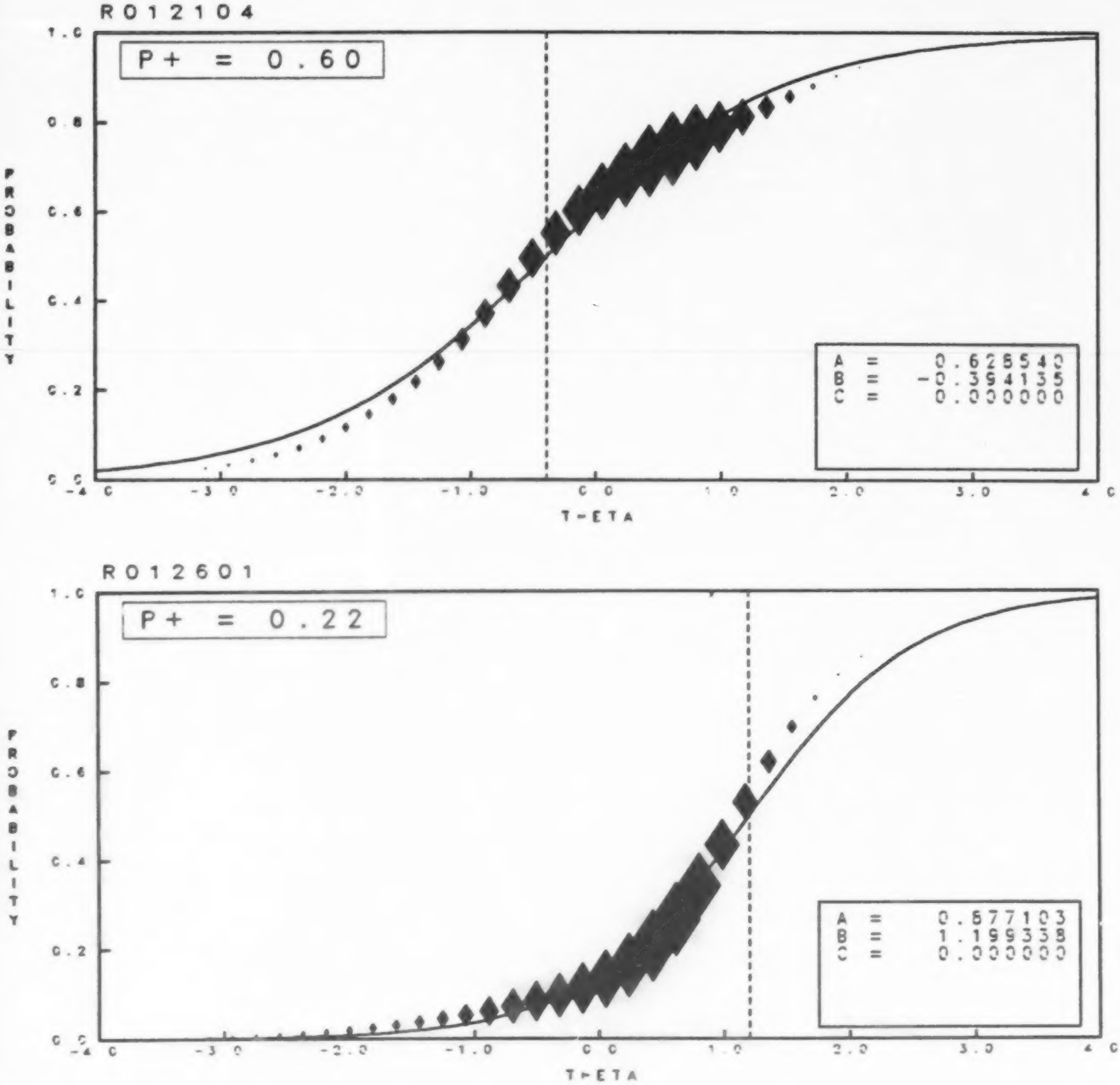
The IRT parameters for the items included in the Trial State Assessment are listed in Appendix D.

9.5 ESTIMATION OF STATE AND SUBGROUP PROFICIENCY DISTRIBUTIONS

The proficiency distributions in each state (and for important subgroups within each state) were estimated by using the multivariate plausible values methodology and the corresponding MGROUP computer program (described in Chapter 8; see also Mislevy, 1991). The MGROUP program (Sheehan, 1985; Rogers, 1991), which was originally based on the procedures described by Mislevy and Sheehan (1987), was used in the 1990 Trial State Assessment of mathematics. The 1992 Trial State Assessment used an enhanced version of MGROUP, based on modifications described by Thomas (1992), to estimate the fourth-grade proficiency distribution for each state. As described in the previous chapter, MGROUP

Figure 9-6

Plots* Comparing Empirical and Model-based Estimates of Item Response Functions for Binary-scored Items Exhibiting Some Model Misfit

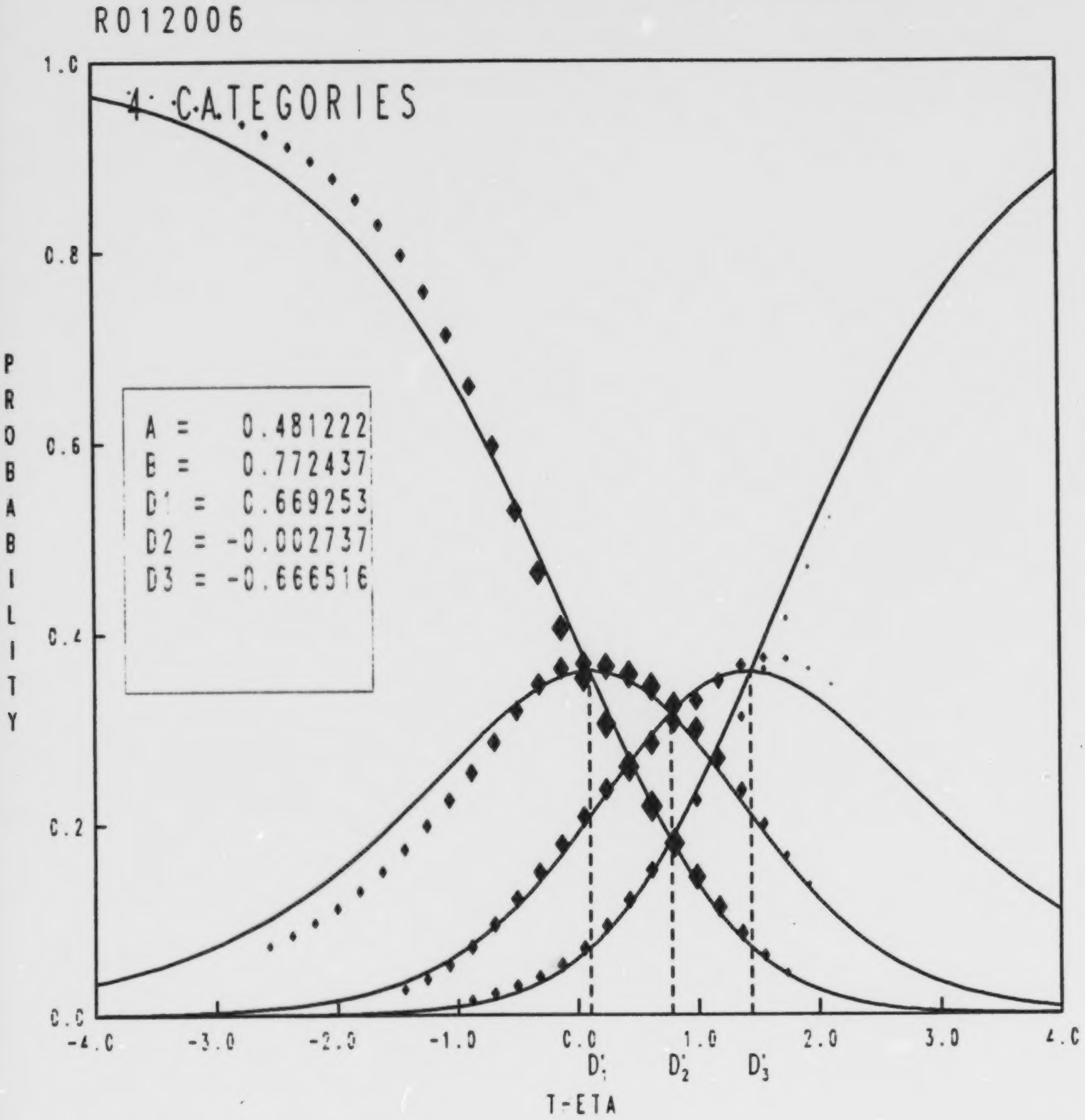


---

* Diamonds indicate estimated conditional probabilities obtained without assuming a logistic form; solid curve indicates estimated item response function assuming a logistic form.

Figure 9-7

Plot* Comparing Empirical and Model-based Estimates of Item Category Characteristic Curves for a Polytomously Scored Item Exhibiting Some Model Misfit

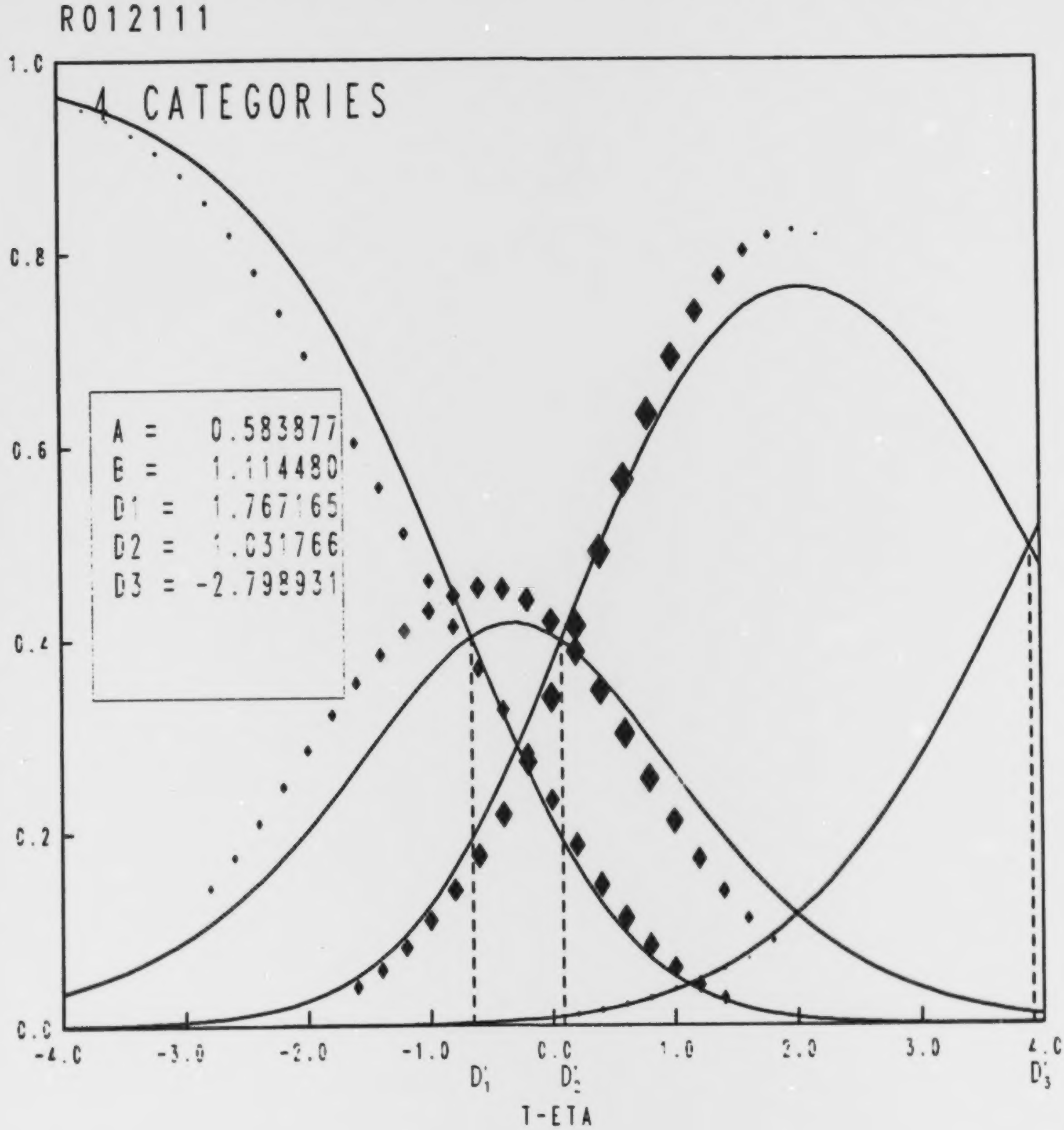


---

* Diamonds indicate estimated conditional probabilities obtained without assuming a logistic form; solid curve indicates estimated item response function assuming a logistic form.

Figure 9-8

Plot* Comparing Empirical and Model-based Estimates of the Item Response Function for Item R012111 Before Collapsing Unsatisfactory and Partial Response Categories

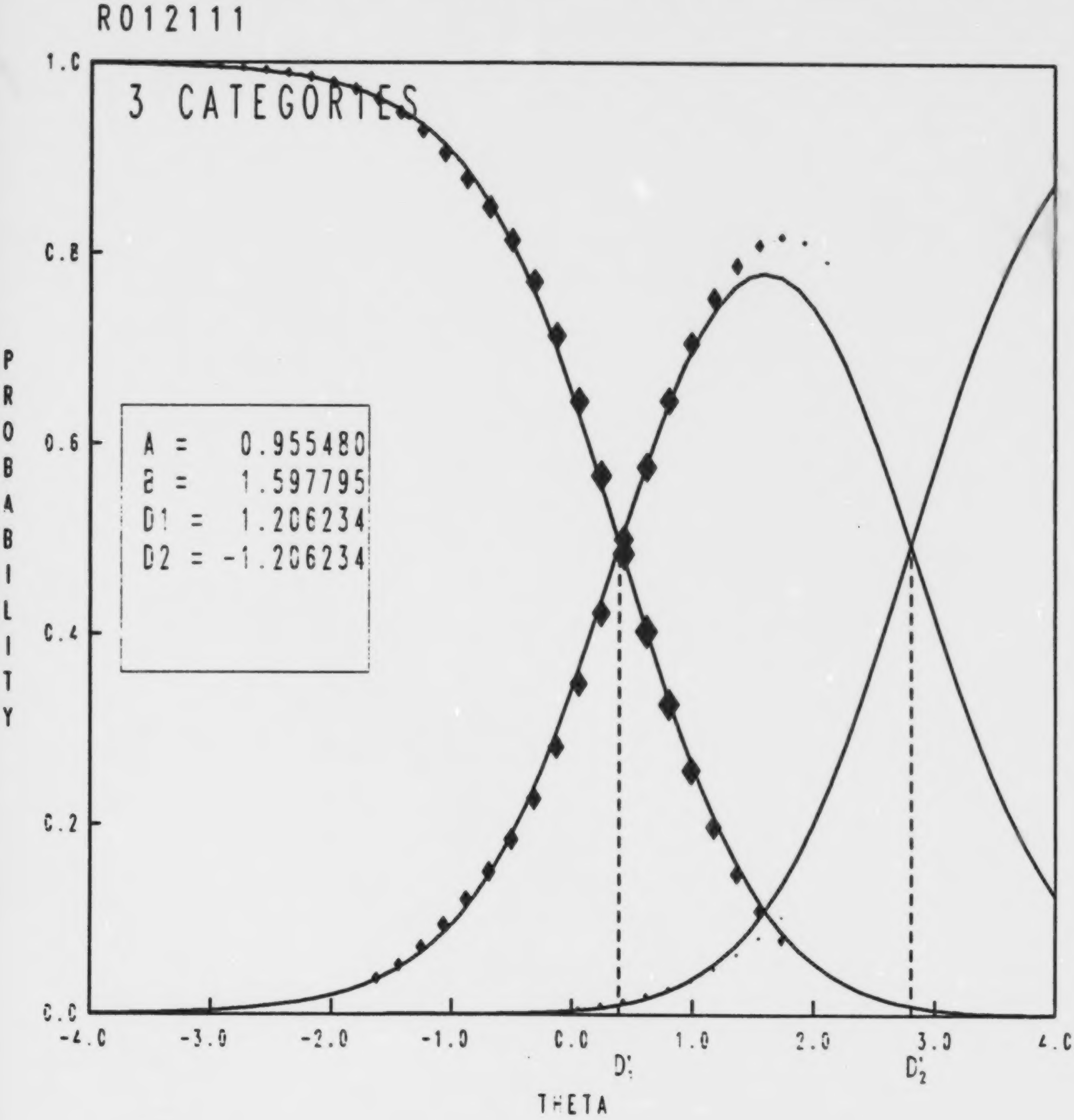


---

* Diamonds indicate estimated conditional probabilities obtained without assuming a logistic form; solid curve indicates estimated item response function assuming a logistic form.

Figure 9-9

Plot Comparing Empirical and Model-based Estimates of the Item Response Function for Item R012111 After Collapsing Unsatisfactory and Partial Response Categories



---

* Diamonds indicate estimated conditional probabilities obtained without assuming a logistic form; solid curve indicates estimated item response function assuming a logistic form.

BEST COPY AVAILABLE

estimates proficiency distributions using information from student's item responses, student background variables, and the item parameter estimates obtained from the BILOG/PARSCALE program.

For the reasons discussed in Mazzeo (1991), separate conditioning models were estimated for each jurisdiction. This resulted in the estimation of 44 distinct conditioning models. The background variables included in each jurisdiction's model (denoted *y* in Chapter 8) were principal component scores derived from the within-state correlation matrix of selected main-effects and two-way interactions associated with a wide range of student, teacher, school, and community variables. A set of five multivariate plausible values was drawn for each student who participated in the Trial State Assessment in reading.

As was the case in the mathematics assessments, plans for reporting each jurisdiction's results required analyses examining the relationships between proficiencies and a large number of background variables. The background variables included student demographic characteristics (e.g., the race/ethnicity of the student, highest level of education attained by parents), students' perceptions about reading, student behavior both in and out of school (e.g., amount of television watched daily, amount of homework done each day), and a variety of other aspects of the students' background and preparation, the background and preparation of their teachers, and the educational, social, and financial environment of the schools they attended.

As described in the previous chapter, to avoid biases in reporting results and to minimize biases in secondary analyses, it is desirable to incorporate measures of a large number of independent variables in the conditioning model. When expressed in terms of contrast-coded main effects and interactions, the number of variables to be included totaled 361. Appendix C provides a listing of the full set of contrasts defined. These contrasts were the common starting point in the development of the conditioning models for each of the participating jurisdictions.

Because of the large number of these contrasts and the fact that, within each jurisdiction, some contrasts had zero variance, some involved relatively small numbers of individuals, and some were highly correlated with other contrasts or sets of contrasts, an effort was made to reduce the dimensionality of the predictor variables in each jurisdiction's MGROUP models. As was done for the 1990 and 1992 Trial State Assessments in mathematics, the original background variable contrasts were standardized and transformed into a set of linearly independent variables by extracting separate sets of principal components (one set for each of the 44 jurisdictions) from the within-jurisdiction correlation matrices of the original contrast variables. The principal components, rather than the original variables, were used as the independent variables in the conditioning [illegible]. As was done for the mathematics assessment, the number of principal components included for each state was the number required to account for approximately 90 percent of the variance in the original contrast variables. Research based on data from the 1990 Trial State Assessment in mathematics suggests that results obtained using such a subset of the components will differ only slightly from those obtained using the full set (Mazzeo, Johnson, Bowker, & Fong, 1992).

Table 9-6 contains a listing of the number of principal components included in and the proportion of proficiency variance accounted for by the conditioning model for each of the 44 participating jurisdictions. It is important to note that the proportion of variance accounted for

Table 9-6

Summary Statistics for State Conditioning Models

State	Number of Principal Components	Proportion of Proficiency Variance Accounted for by the Reading for Literary Experience Scale	Proportion of Proficiency Variance Accounted for by the Reading to Gain Information Scale	Conditional Correlation Between Scales
Alabama	150	0.59	0.59	.80
Arizona	162	0.50	0.53	.93
Arkansas	157	0.54	0.57	.89
California	156	0.63	0.65	.84
Colorado	159	0.51	0.56	.91
Connecticut	153	0.58	0.65	.82
Delaware	153	0.61	0.66	.94
District of Columbia	146	0.58	0.59	.81
Florida	162	0.53	0.55	.89
Georgia	159	0.53	0.55	.92
Guam	108	0.57	0.58	.78
Hawaii	162	0.54	0.52	.84
Idaho	161	0.50	0.50	.81
Indiana	151	0.47	0.47	.92
Iowa	152	0.46	0.51	.84
Kentucky	155	0.52	0.52	.90
Louisiana	156	0.49	0.51	.81
Maine	143	0.49	0.49	.77
Maryland	155	0.60	0.60	.87
Massachusetts	154	0.56	0.60	.84
Michigan	148	0.57	0.59	.84
Minnesota	144	0.51	0.52	.84
Mississippi	154	0.52	0.53	.91
Missouri	154	0.52	0.55	.84
Nebraska	145	0.50	0.52	.89
New Hampshire	152	0.53	0.49	.93
New Jersey	152	0.67	0.65	.82
New Mexico	149	0.54	0.54	.93
New York	155	0.61	0.63	.89
North Carolina	161	0.54	0.56	.87
North Dakota	140	0.48	0.56	.92
Ohio	151	0.49	0.54	.87
Oklahoma	154	0.49	0.53	.92
Pennsylvania	156	0.53	0.60	.88
Rhode Island	146	0.58	0.63	.80
South Carolina	162	0.56	0.58	.87
Tennessee	158	0.51	0.56	.80
Texas	158	0.55	0.60	.87
Utah	161	0.49	0.51	.88
Virginia	160	0.58	0.57	.88
Virgin Islands	99	0.67	0.73	.73
West Virginia	152	0.50	0.56	.84
Wisconsin	155	0.54	0.51	.79
Wyoming	155	0.53	0.52	.81

by the conditioning model differs across scales within a state, and across states within a scale. Such variability is not unexpected for at least two reasons. First, there is no reason to expect the strength of the relationship between proficiency and demographics to be identical across all states. In fact, one of the reasons for fitting separate conditioning models is that the strength and nature of this relationship may differ across states. Second, the homogeneity of the demographic profile also differs across states. As with any correlational analysis, the restriction of the range in the predictor variables will attenuate the relationship.

Table 9-6 also provides the estimated within-jurisdiction correlation between the two scales. The values, taken directly from the revised MGROUP program, are estimates of the within-jurisdiction correlations *conditional on the set of principal components included in the conditioning model*. The number and nature of the scales that were produced were consistent with the recommendations for reporting that were given by the National Assessment Planning Project (see Chapter 2). Reporting results on multiple scales is typically most informative when each of the scales provides unique information about the profile of knowledge and skills possessed by the students being assessed. In such cases, one would hope to see relatively low correlations among the subscales. However, with a couple of exceptions, the correlations between the scales are high across all jurisdictions, always exceeding .7 and quite often exceeding .9. This is particularly noteworthy when one considers that these are correlations *conditional* on a rather large set of background variables. The *marginal* correlations between content area scales would be higher, particularly for those correlations in the .7 to .8 range.

As discussed in Chapter 8, NAEP scales are viewed as summaries of consistencies and regularities that are present in item-level data. Such summaries should agree with other reasonable summaries of the item-level data. In order to evaluate the reasonableness of the scaling and estimation results, a variety of analyses were conducted to compare state-level and subgroup-level performance in terms of the content area scaled scores and in terms of the average proportion correct for the set of items in a content area. High agreement was found in all of these analyses. One set of such analyses is presented in Figure 9-10. The figure contains scatterplots of the state item score mean versus the state scale score means, for each of the two reading content areas and the composite scale. As is evident from the figures, there is an extremely strong relationship between the estimates of state-level performance in the scale-score and item-score metrics for both content areas.

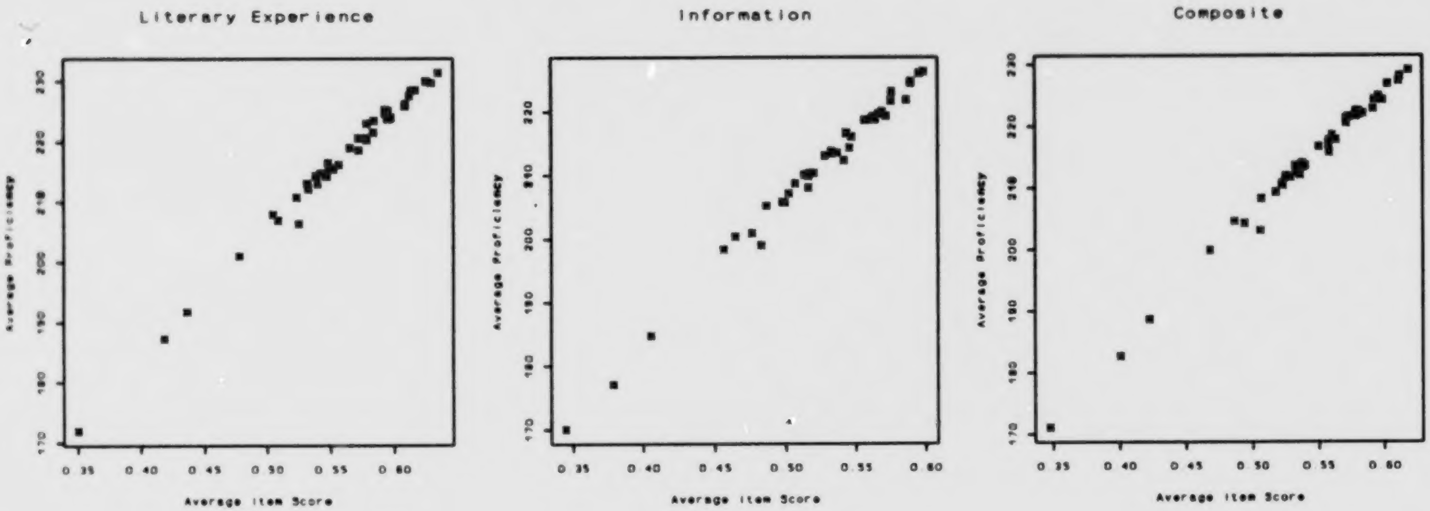
9.6 LINKING STATE AND NATIONAL SCALES

A major purpose of the Trial State Assessment Program was to allow each participating jurisdiction to compare its 1992 results with the nation as a whole and with the region of the country in which that jurisdiction is located. For meaningful comparisons to be made between each of the Trial State Assessment jurisdictions and the relevant national sample, results from these two assessments had to be expressed in terms of a similar system of scale units.

The purpose of this section is to describe the procedures used to align the 1992 Trial State scales with their 1992 national counterparts. The procedures that were used are similar to the common population equating procedures employed to link the 1990 national and state mathematics scales (Mazzeo, 1991; Yamamoto & Mazzeo, 1992).

Figure 9-10

Plot of Mean Proficiency Versus Mean Item Score



Literary Experience

Information

Composite

Using the sampling weights provided by Westat, the combined sample of students from participating jurisdictions (a total sample size of 107,689) was used to estimate the distribution of proficiencies for the population of students enrolled in public schools in the participating states and the District of Columbia[5]. Data were also used from a subsample of 4,618 students in the national assessment at grade 4, consisting of grade-eligible public-school students from 42 of the 44 jurisdictions who participated in the 1992 Trial State Assessment. Appropriate weights were provided by Westat to obtain estimates of the distribution of proficiency for the same target population.

Thus, for each of the two scales, two sets of proficiency distributions were obtained. One set, based on the sample of combined data from the Trial State Assessment (referred to as the Trial State Assessment Aggregate Sample) and using item parameter estimates and conditioning results from that assessment, was in the metric of the 1992 Trial State Assessment. The other, based on the sample from the 1992 national assessment (referred to as the State Aggregate Comparison, or SAC, sample) and obtained using item parameters and conditioning results from that assessment, was in the metric of the 1992 national assessment. The latter metric had already been set using procedures described in the technical report of the 1992 national assessment. The two Trial State Assessment and national scales were made comparable by constraining the mean and standard deviation of the two sets of estimates to be equal.

More specifically, the following steps were followed to linearly link the scales of the two assessments:

1) For each scale, estimates of the proficiency distribution for the Trial State Assessment Aggregate Sample were obtained using the full set of plausible values generated by the MGROUP program. The weights used were the final sampling weights provided by Westat, not the rescaled versions discussed in section 9.3. For each scale, the arithmetic mean of the five sets of plausible values was taken as the overall estimated mean and the geometric mean of the standard deviations of the five sets of plausible values was taken as the overall estimated standard deviation.

2) For each scale, the estimated proficiency distribution of the State Aggregate Comparison sample was obtained, again using the full set of plausible values generated by the MGROUP program. The weights used were specially provided by Westat to allow for the estimation of proficiency for the same target population of students estimated by the state data. The means and standard deviations of the distributions for each scale were obtained for this sample in the same manner as described in step 1. These means and standard deviations were then linearly adjusted to reflect the reporting metric used for the national assessment (see the technical report for the NAEP 1992 national assessment.

---

[5] Students from Guam and the Virgin Islands were excluded from the definition of this target population; hence, data from students from these jurisdictions were not included in the combined Trial State Assessment samples.

3) For each scale, a set of linear transformation coefficients were obtained to link the state scale to the corresponding national scale. The linking was of the form

$$Y^* = k_1 + k_2 Y$$

where

Y = a scale level in terms of the system of units of the provisional BILOG/PARSCALE scale of the Trial State Assessment scaling

Y^* = a scale level in terms of the system of units comparable to those used for reporting the 1992 national reading results

k_2 = $[\text{Standard-Deviation}_{SAC}]/[\text{Standard-Deviation}_{TSA}]$

k_1 = $\text{Mean}_{SAC} - k_2[\text{Mean}_{TSA}]$

The final conversion parameters for transforming plausible values from the provisional BILOG/PARSCALE scales to the final Trial State Assessment reporting scales are given in Table 9-7. All Trial State Assessment results are reported in terms of the Y^* metric.

Table 9-7
Transformation Constants

Scale	k_1	k_2
Reading for Literary Experience	217.56	38.10
Reading to Gain Information	212.50	37.00

As evident from the discussion above, a linear method was used to link the scales from the Trial State and national assessments. While these linear methods ensure equality of means and standard deviations for the Trial State Assessment aggregate (after transformation) and the SAC samples, they do not guarantee the shapes of the estimated proficiency distributions for the two samples to be the same. As these two samples are both from a common target population, estimates of the proficiency distribution of that target population based on each of the samples should be quite similar in shape in order to justify strong claims of comparability for the Trial State and national scales. Substantial differences in the shapes of the two estimated distributions would result in differing estimates of the percentages of students above achievement levels or of percentile locations depending on whether Trial State or national scales were used—a clearly unacceptable result given claims about comparability of scales. In the face of such results, nonlinear linking methods would be required.

Analyses were carried out to verify the degree to which the linear linking process described above produced comparable scales for Trial State and national results. Comparisons were made between two estimated proficiency distributions, one based on the Trial State Assessment aggregate and one based on the SAC sample, for each of the two reading scales.

The comparisons were carried out using slightly modified versions of what Wainer (1974) refers to as suspended rootograms. The final reporting scales for the Trial State and national assessments were each divided into 10-point intervals. Two sets of estimates of the percentage of students in each interval were obtained, one based on the Trial State Assessment aggregate sample and one based on the SAC sample. Following Tukey (1977), the square root of these estimated percentages were compared.[6]

The comparisons are shown in Figure 9-11. The heights of each of the unshaded bars correspond to the square root of the percentage of students from the Trial State Assessment aggregate sample in each 10-point interval on the final reporting scale. The shaded bars show the differences in root percents between the Trial State Assessment and SAC estimates[7]. Positive differences indicate intervals in which the estimated percentages from the State Aggregate Comparison sample are lower than those obtained from the Trial State Assessment aggregate. Conversely, negative differences indicate intervals in which the estimated percentages from the State Aggregate Comparison sample are higher. For both scales, differences in root percents are quite small, suggesting that the shapes of the two estimated distributions are quite similar (i.e., unimodal with slight negative skewness). There is some evidence that the estimates produced using the Trial State Assessment data are slightly heavier in the extreme lower tails (below 100). However, even these differences at the extremes are small in magnitude and have little impact on estimates of reported statistics such as percentages of students below the achievement levels.

9.7 PRODUCING A READING COMPOSITE SCALE

For the national assessment, a composite scale was created for the fourth grade as an overall measure of reading proficiency. The composite was a weighted average of plausible values on the two content area scales (Reading for Literary Experience and Reading to Gain Information). The weights for the national content area scales were proportional to the relative importance assigned to each content area for the fourth grade in the assessment specifications developed by the Reading Objectives Panel. Consequently, the weights for each of the content areas are similar to the actual proportion of items from that content area.

The Trial State Assessment composite scale was developed using weights identical to those used to produce the composites for the 1992 national reading assessment. The weights are given in Table 9-8. In developing the Trial State Assessment composite the weights were

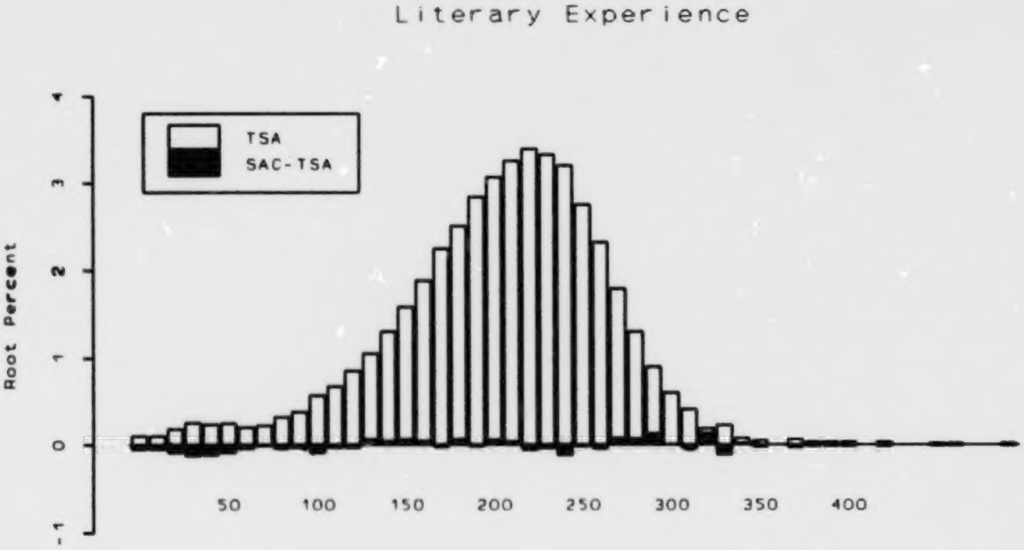
---

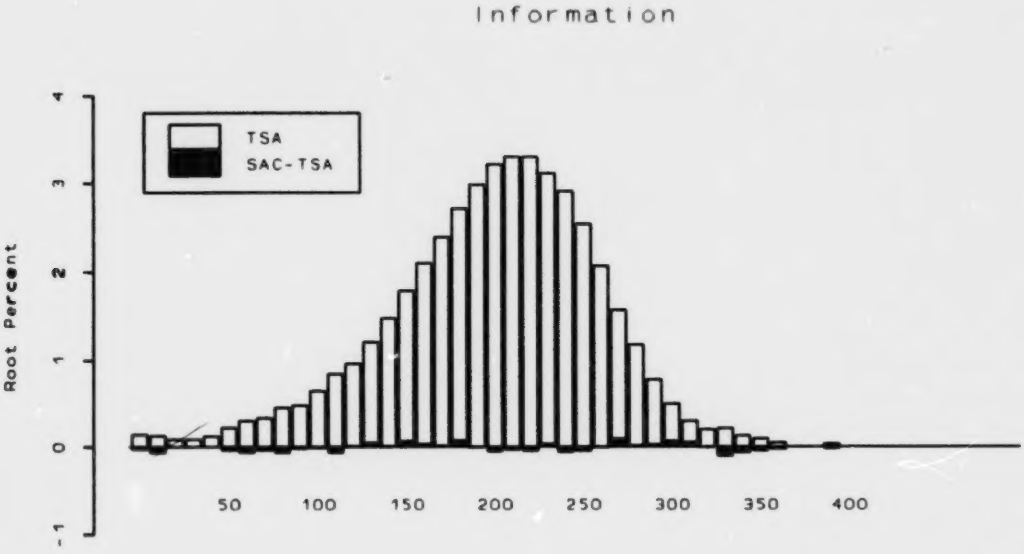
[6] The square root transformation allows for more effective comparisons for counts (or equivalently, percentages) when the expected number of counts in each interval is likely to vary greatly over the range of intervals, as is the case for the NAEP scales where the expected counts of individuals in intervals near the extremes of the scale (e.g., below 150 and above 350) are dramatically smaller than the counts obtained near the middle of the scale.

[7] Wainer (1974), among others, has suggested that looking at residuals around a fitted straight line makes judgments of differences somewhat easier to make. Hence, the *differences between the root percents*—rather than separate sets of root percents—from the SAC sample and the Trial State Assessment aggregate are plotted around the x-axis in Figures 9-11 and 9-12.

Figure 9-11

Rootogram Comparing Proficiency Distributions for the Trial State Assessment Aggregate Sample and the State Aggregate Comparison Sample from the National Assessment for Each Content Area Scale





applied to the plausible values for each content area scale as expressed in terms of the final Trial State Assessment scales (i.e., after transformation from the provisional BILOG/PARSCALE scales.)

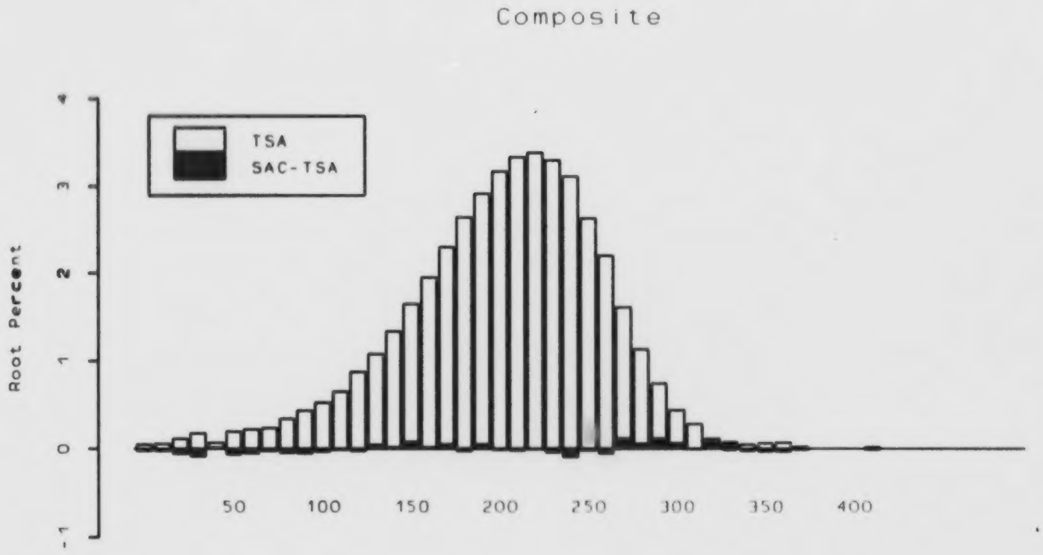
Table 9-8
Weights Used for Each Scale to Form the Reading Composite

Scale	Weights
Reading for Literary Experience	.55
Reading to Gain Information	.45

Figure 9-12 provides a rootogram comparing the estimated proficiency distributions based on the Trial State Assessment and SAC samples for the grade 4 composite. Consistent with the results presented separately by scale, there is some evidence that the estimates produced using the Trial State Assessment data are slightly heavier in the lower tail than the corresponding estimate based on the SAC data. Again, however, the differences in root relative percents are small in magnitude.

Figure 9-12

Rootogram Comparing Proficiency Distributions
for the Trial State Assessment Aggregate Sample
and the State Aggregate Comparison Sample from the National Assessment
for the Composite Scale



Chapter 10

CONVENTIONS USED IN REPORTING THE RESULTS OF THE 1992 TRIAL STATE ASSESSMENT IN READING

John Mazzeo

Educational Testing Service

10.1 OVERVIEW

The primary results of the 1992 Trial State Assessment in reading were released as four separate reports: a *Reading Report* for each state, the *NAEP 1992 Reading Report Card for the Nation and the States*, the *Data Compendium from the NAEP 1992 Reading Assessment for the Nation and the States*, and a six-section almanac of data for each state.

The *Reading Report* is a computer-generated report that provides, for each state, reading results for their fourth-grade students. While national and regional results[1] are included for comparison purposes, the major focus of each of these computer-generated reports is the results for that particular jurisdiction. School and student participation rates are reported for each jurisdiction to provide information about the generalizability of the results. School participation rates are reported both in terms of the initially selected samples of schools and in terms of the finally achieved samples, including replacement schools. Several different student participation rates are reported, including the overall rate, the percentage of students excluded from the assessment, and the exclusion rates for Limited English Proficiency (LEP) students and for students with Individualized Education Plans (IEPs).

The text and tables of each state's *Reading Report* were produced by a computerized report generation system developed by ETS report writers, statisticians, data analysts, graphic designers, and editors. The reports contain state-level estimates of proficiency means and standard deviations, proportions of students at or above achievement levels defined by the National Assessment Governing Board (NAGB) (see Appendix F for details on the definition and development of these levels), proportions of students at or above the traditional NAEP anchor levels, and selected percentiles for the state as a whole and for subgroups defined by four key reporting variables (referred to here as primary reporting variables): 1) gender, 2) race/ethnicity, 3) level of parents' education, and 4) type of community. In addition, proficiency means are also reported for a variety of other subpopulations (referred to as secondary reporting variables) defined by responses to items from the student, teacher, and school questionnaires and by school and community demographic variables provided by Westat[2]. Item-level results are also provided for the two released extended constructed-response items included in the 1992 reading assessment. The NAEP computer-generated reporting system is described in detail by Jerry (1993).

---

[1] The national and regional results included in the state reports and in the portions of the *Data Compendium from the NAEP 1992 Reading Assessment for the Nation and the States* that present state results are based on data from the 1992 national reading assessment and include fourth-grade students enrolled in public schools.

A second report, the *NAEP 1992 Reading Report Card for the Nation and the States*, highlights key assessment results for the nation and summarizes results across the states and territories participating in the assessment. This report contains composite scale results (proficiency means, proportions at or above achievement levels, etc.) for the nation, each of the four regions of the country, and each jurisdiction participating in the Trial State Assessment, both overall and by the primary reporting variables. In addition, overall results are reported for the reading scales. The *Report Card* also contains a number of specially developed graphical displays that summarize and compare results for the full set of Trial State Assessment participants.

The third report is entitled *Data Compendium from the NAEP 1992 Reading Assessment for the Nation and the States*. Like the *Report Card*, the *Compendium* reports results for the nation and for all of the states and territories participating in the Trial State Assessment. However, unlike the *Report Card*, the *Compendium* is primarily tabular in nature and contains little in the way of interpretive text. The *Compendium* contains most of the tables included in the *Report Card* plus additional tables that provide composite scale results for a large number of secondary reporting variables. The variables used to report each jurisdiction's results in the individual state reports are included in the *Compendium*, along with additional background variables derived from the student, teacher, and school questionnaires.

The fourth report is a six-section almanac. The first section, or "distribution" section, provides results for the achievement levels and percentiles. Three of the sections (referred to as proficiency sections) present analyses based on responses to each of the questionnaires (student, reading teacher, and school) administered as part of the Trial State Assessment. For most background questions contained in these questionnaires, the proportion of students responding to each option and the reading composite proficiency mean for these students are reported with their jackknifed standard errors. The student proficiency section of the almanac also contains selected percentiles and the percentages of students at or above achievement levels and anchor points. Results are provided for the total group of students in each participating jurisdiction, as well as for groups defined by the primary reporting variables (gender, race/ethnicity, type of community, and level of parents' education). The fifth section of the almanac, the scale section, reports proficiency means and associated standard errors for the two reading scales. Results in this section are also reported for the total group in each state, as well as for select subgroups of interest. The final section of the almanac, the "p-value" section, provides the total-group proportion of correct responses to each cognitive item included in the assessment.

The production of the state reports, the *Report Card*, the *Data Compendium*, and the almanacs required a large number of decisions about a variety of data analysis and statistical issues. For example, a wide variety of demographic profiles and instructional practices exist across the states and territories that participated in the Trial State Assessment. Given the sample sizes obtained for each state, certain categories of the reporting variables contained limited numbers of examinees. A decision was needed as to what constituted a sufficient sample size to permit the reliable reporting of subgroup results, and which, if any, estimates were sufficiently unreliable to need to be identified (or flagged) as a caution to readers. As a second example, the state report contained computer-generated text that described the results for a particular state and compared total and subgroup performance within the state to that of the region and nation. A number of inferential rules, based on logical and statistical considerations, had to be developed to ensure that the computer-generated reports were coherent from a substantive standpoint and were based on statistical principals of significance testing.

---

[2] Some of these variables were used by Westat in developing the sampling frame for the assessment and in drawing the sample of participating schools.

The purpose of this chapter is to document the major conventions and statistical procedures used in generating the state reports, the *Report Card*, the *Data Compendium*, and the almanacs. The principal focus of this chapter is on conventions used in the production of the computer-generated state reports. However, sections 10.2 to 10.4 contain material applicable to all four summary reports. Additional details about procedures relevant to the *Report Card* and *Data Compendium* can be found in the text and technical appendices of those reports.

10.2 MINIMUM SAMPLE SIZES FOR REPORTING SUBGROUP RESULTS

In all four reports, estimates of quantities such as composite and content area proficiency means, percentages of students at or above the achievement levels, and percentages of students indicating particular levels of background variables (as measured in the student, teacher, and school questionnaires) are reported for the total population of fourth-grade students in each jurisdiction, as well as for certain key subgroups of interest. The subgroups were defined by four primary NAEP reporting variables. NAEP reports results for five racial/ethnic subgroups (White, Black, Hispanic, Asian American/Pacific Islander, and American Indian/Alaskan Native), four types of communities (advantaged urban, disadvantaged urban, extreme rural, and other non-extreme communities), four levels of parents' education (did not finish high school, high school graduate, some college, college graduate), and by gender (males, females). However, in some jurisdictions, and for some regions of the country, sample sizes were not large enough to permit accurate estimation of proficiency and/or background variable results for one or more of the categories of these variables.

For results to be reported for any category, a minimum sample size of 62 was required. This number was arrived at by determining the sample size required to detect an effect size of 0.5 with a probability of .8 or greater[3]. The effect size of 0.5 pertains to the "true" difference in mean proficiency between the subgroup in question and the total fourth-grade public-school population in the state, divided by the standard deviation of proficiency in the total population. The same convention was used in reporting the 1990 and 1992 Trial State Assessment results in mathematics.

---

[3] A design effect of 2 was assumed for this purpose, implying a sample design-based variance twice that of simple random sampling. This is consistent with previous NAEP experience (Johnson & Rust, 1992).

The summary reports also include large numbers of tables that provide estimates of the proportion of the students responding to each category of a secondary reporting variable, as well as the mean proficiency of the students within each category. In several instances, the number of students in a particular category of these background variables was also less than 62. The same minimum sample size restriction of 62 was applied to these subgroups as well.

10.3 ESTIMATES OF STANDARD ERRORS WITH LARGE MEAN SQUARED ERRORS

Standard errors of mean proficiencies, proportions, and percentiles play an important role in interpreting subgroup results and comparing the performances of two or more subgroups. The jackknife standard errors reported by NAEP are statistics whose quality depends on certain features of the sample from which the estimate is obtained. In certain cases, typically when the number of students upon which the standard error is based is small or when this group of students all come from a small number of participating schools, the mean squared error associated with the estimated standard errors may be quite large. In the summary reports, estimated standard errors subject to large mean squared errors are followed by the symbol "!".

The magnitude of the mean squared error associated with an estimated standard error for the mean or proportion of a group depends on the coefficient of variation (*CV*) of the estimated size of the population group, denoted as *N*. This coefficient of variation is estimated by:

$$CV(\hat{N}) = \frac{SE(\hat{N})}{\hat{N}}$$

where $\hat{N}$ is a point estimate of N and $SE(\hat{N})$ is the jackknife standard error of $\hat{N}$.

Experience with previous NAEP assessments suggests that when this coefficient exceeds 0.2, the mean squared error of the estimated standard errors of means and proportions based on samples of this size may be quite large. Therefore, the standard errors of means and proportions for all subgroups for which the coefficient of variation of the population size exceeds 0.2 are followed by "!" in the tables of all summary reports. These standard errors, and any confidence intervals or significance tests involving these standard errors, should be interpreted with caution. (Further discussion of this issue can be found in Johnson & Rust, 1992.)

10.4 TREATMENT OF MISSING DATA FROM THE STUDENT, TEACHER, AND SCHOOL QUESTIONNAIRES

Responses to the student, teacher, and school questionnaires played a prominent role in all reports. Although the return rates on all three types of questionnaire were high[4], there were missing data from each type.

For the questionnaires, the reported estimated percentages of students in the various categories of background variables, and the estimates of the mean proficiency of such groups, were based on only those students for whom data on the background variable were available. The analyses pertaining to a particular background variable presented in the state reports and the *Data Compendium* assumed the data were missing completely at random in the sense that the mechanism generating the missing data was assumed to be independent of both the response to the particular background variables and to proficiency.

The estimates of proportions and proficiencies based on the "missing-completely-at-random" assumption are possibly subject to nonresponse bias if the assumption is not correct. The amount of missing data was small (usually, less than 2 percent) for most of the variables obtained from the student and school questionnaires. For analyses based on these variables, reported results are subject to little, if any, nonresponse bias. However, for particular background questions from the student and school questionnaires, the level of nonresponse in certain jurisdictions was somewhat higher. As a result, the possibility for nonresponse bias in the results for this latter set of questions is also somewhat greater. Background questions for which more than 10 percent of the returned questionnaires were missing are identified in background almanacs produced for each jurisdiction. Again, results for analyses involving these questions should be interpreted with caution.

In order to analyze the relationships between teachers' questionnaire responses and their students' achievement, each teacher's questionnaire had to be matched to all of the students who were taught mathematics by that teacher. Table 10-1 provides percentages of fourth-grade students that were matched to teacher questionnaires for each of the 44 jurisdictions that participated in the Trial State Assessment. The percentages were calculated using the final sampling weights provided by Westat (see Chapters 7 and 9).

Three separate match rates are given in the table. The first is the percentage of students that could not be matched to either part of the two-part teacher questionnaire. The second match rate is the percentage of students that could be matched to only the first part of the teacher questionnaire. The third is the percentage of students that could be matched to both the first and second parts of the teacher questionnaire. Note that these match rates do not reflect the additional missing data due to item-level nonresponse. The amount of additional item-level nonresponse in the returned teacher questionnaires can also be found in the almanacs produced for each jurisdiction.

---

[4] Information about survey participation rates (both school and student), as well as proportions of students excluded by each state from the assessment, are given in Appendix B. Adjustments intended to account for school and student nonresponse are described in Chapter 8.

Table 10-1
Weighted Percentage of Students Matched to Reading Teacher Questionnaires

State	Questionnaire Match Rate		
	No Match	Part I Only	Parts I and II
Alabama	4.3	95.7	90.4
Arizona	3.8	96.2	90.1
Arkansas	2.7	97.3	93.2
California	4.4	95.6	88.9
Colorado	6.9	93.1	82.2
Connecticut	5.7	94.3	87.5
Delaware	0.7	99.3	95.5
District of Columbia	11.2	88.8	73.8
Florida	6.9	93.1	88.4
Georgia	6.3	93.7	86.9
Guam	2.6	97.4	92.5
Hawaii	2.3	97.7	92.2
Idaho	0.6	99.4	93.3
Indiana	2.8	97.2	89.2
Iowa	4.1	95.9	89.8
Kentucky	3.9	96.1	90.0
Louisiana	4.8	95.2	87.5
Maine	6.7	93.3	85.8
Maryland	4.8	95.2	90.0
Massachusetts	4.4	95.6	88.1
Michigan	3.7	96.3	87.2
Minnesota	9.3	90.7	74.7
Mississippi	4.7	95.3	86.7
Missouri	1.6	98.4	89.7
Nebraska	1.9	98.1	82.1
New Hampshire	3.5	96.5	93.7
New Jersey	1.8	98.2	92.1
New Mexico	7.1	92.9	81.1
New York	4.4	95.6	88.8
North Carolina	4.2	95.8	89.5
North Dakota	1.6	98.4	90.4
Ohio	5.2	94.8	86.9
Oklahoma	1.2	98.8	91.5
Pennsylvania	1.6	98.4	90.7
Rhode Island	3.5	96.5	88.2
South Carolina	0.8	99.2	94.2
Tennessee	3.0	97.0	91.2
Texas	5.7	94.3	85.3
Utah	1.6	98.4	91.6
Virginia	5.1	94.9	91.1
Virgin Islands	7.1	92.9	76.5
West Virginia	4.4	95.6	87.1
Wisconsin	2.8	97.2	89.5
Wyoming	4.8	95.2	85.4

10.5 STATISTICAL RULES USED FOR PRODUCING THE STATE REPORTS

As described earlier, the state reports contain state-level estimates of fourth-grade mean proficiencies, proportions of students at or above selected scale points, and percentiles for the state as a whole and for the categories of a large number of reporting variables. Similar results are provided for the nation and, where sample sizes permitted, for the region to which each state belongs[5]. The state reports were computer-generated. The tables and figures, as well as the text of the report, were automatically tailored for each jurisdiction based on the pattern of results obtained. The purpose of this section is to describe some of the procedures and rules used to produce these individually tailored reports. A more detailed presentation is given by Jerry (1993).

In the 1992 state reports, the results are presented principally through figures and tables containing estimated means, proportions, and percentiles, along with their standard errors. In addition to the figures and tables, computer-generated interpretive text is also provided. In a large number of cases, the computer-generated interpretive text is primarily descriptive in nature and reports the total group and subgroup proficiency means and proportions of interest. However, some of the interpretive text focuses on interesting and potentially important group differences in reading proficiency or on the percentages of students responding in particular ways to the background questions. Additional interpretive text compares state-level results with those of the nation. Rules were developed to produce the computer-generated text for questions involving the comparison of results for subgroups and interpretations of patterns of results. These rules were based on a variety of considerations, including a desire for 1) statistical rigor in the identification of important group differences and patterns of results, and 2) solutions that were within the limitations imposed by the availability of computational resources and the time frame for the production of the report. The following sections describe some of these procedures and rules.

10.5.1 Comparing Means and Proportions for Mutually Exclusive Groups of Students

Many of the group comparisons explicitly commented on in the state reports involved mutually exclusive sets of students. One common example of such a comparison is the contrast between the mean composite proficiency in a particular state and the mean composite proficiency in the nation. Other examples include comparisons within a jurisdiction of the average proficiency for male and female students; White and Hispanic students; students from advantaged urban schools and disadvantaged urban schools; and students who reported watching six or more hours of television each night and students who report watching less than one hour each night.

In the state reports, computer-generated text indicated that means or proportions from two groups were different only when the difference in the point estimates for the groups being compared was statistically significant based on a two-sided test carried out at an approximate α level of .05. A large-sample procedure was used for determining statistical significance that

---

[5] Because United States territories are not classified into NAEP regions, no regional comparisons were provided for Guam.

NAEP staff felt was reasonable from a statistical standpoint, as well as being computationally tractable. The procedure was as follows.

Let t_i be the statistic in question (i.e., a mean or proportion for group i) and let $SE(t_i)$ be the jackknife standard error of the statistic. The computer-generated text in the state report identified the means or proportions for groups i and j as being different if and only if:

$$\frac{|t_i - t_j|}{\sqrt{\hat{SE}^2(t_i) + \hat{SE}^2(t_j)}} \geq Z_{\alpha^*}$$

where α^* was defined as .05/2, and Z_{α^*} is the $(1 - \alpha^*)$ percentile of the standard normal distribution. In cases where group comparisons were treated as individual units (for example, comparing the results obtained by males to those obtained by females), the test statistic is equivalent to a standard two-tailed t-test for the difference between group means or proportions from large independent samples with the significance (α) level set at .05.

The large-sample procedures described in this section assume that the data being compared are from independent samples. Because of the sampling design used for the Trial State Assessment, in which both schools and students within schools are randomly sampled, the data from mutually exclusive sets of students within a state may not be strictly independent. Therefore, the significance tests employed are, in many cases, only approximate. Another procedure, one that does not assume independence, could have been conducted. However, that procedure is computationally burdensome and resources precluded its application for all the comparisons in the state reports. It was the judgment of NAEP staff that if the data were correlated across groups, in most cases the correlation was likely to be positive. Since, in such instances, significance tests based on assumptions of independent samples are conservative (because the estimated standard error of the difference based on independence assumptions is larger than the more complicated estimate based on positively correlated groups), the approximate procedure was used for all comparisons presented in the *State Reports*.

The procedures described above were also used for testing differences of both means *and* proportions. The approximation for the test for proportions works best when sample sizes are large, and the proportions being tested have magnitude close to .5. Statements about group differences should be interpreted with caution if at least one of the groups being compared is small in size and/or if somewhat extreme proportions are being compared.

10.5.2 Multiple Comparison Procedures

Frequently, groups (or families) of comparisons were made and were presented as a single set. The appropriate text, usually a set of sentences or a paragraph, was selected for inclusion in the report based on the pattern of results for the entire set of comparisons. For example, in Chapter 1 of the state report, state/territory results were compared to national results for both of the reading scales. For families of contrasts like these, a Bonferroni procedure was used for determining the value of Z_{α^*} in the equation given in the previous section. Specifically, in the case of multiple group comparisons, where α^* was defined as $.05/2c$, and c was the number of contrasts in the set . In this example, c was taken to be 2, and each statistical test was consequently carried out at a two-tailed significance level of .05/2. As a second example, Chapter 2 of the state report contained a section that compared average proficiencies for a majority group (in the case of race/ethnicity, for example, usually White students) to those obtained by each minority group containing 62 or more students. Assuming three such minority groups, the text in the section was based on the results of three predefined statistical tests (i.e., a test comparing majority group performance to that of each of the three minority groups). Each statistical test was carried out at a two-tailed significance level of .05/3.

10.5.3 Determining the Highest and Lowest Scoring Groups from a Set of Ranked Groups

Certain analyses in the state report consisted of determining which of a set of several groups had the highest or lowest proficiency among the set. For example, one analysis compared the average proficiency of students who reported reading various numbers of books outside of school during the past month. There were four levels of book reading—none, one or two, three or four, and five or more. Based on their answers to this question in the student background questionnaire, students were classified into one of the four levels of book reading, and the mean composite proficiency was obtained for students at each level. The analysis focused on which, if any, of the groups had the highest and lowest mean composite proficiency.

The analyses were carried out using the statistics described in the previous section. The groups were ranked from highest to lowest in terms of their estimated mean proficiency. Then, three separate significance tests were carried out: 1) the highest group was compared to the lowest group; 2) the highest group was compared to the second highest group; and 3) the lowest group was compared to the second lowest group. The following conclusions were drawn:

- If all three comparisons were statistically significant, the performance of the highest ranking group was described as *highest* and the performance of the lowest ranking group was described as *lowest.*

- If only the first and second tests were significant, the highest ranking group was described as *highest*, but no comment was made about the lowest ranking group.

- Similarly, if only the first and third tests were significant, the lowest ranking group was described as *lowest*, but no comment was made about the highest ranking group.

- If only the first test was significant, the highest group was described as performing better than the lowest group, but no *highest* and *lowest* group were designated.

The Bonferroni adjustment factor was taken as the number of possible pairwise comparisons because of the ranking of groups prior to the carrying out of significance tests.

10.5.4 Statistical Significance and Estimated Effect Sizes

Whenever single comparisons were made between groups, an attempt was made to distinguish between group differences that were statistically significant but rather small in a practical sense and differences that were both statistically and practically significant. In order to make such distinctions, a procedure based on estimated effect sizes was used. The estimated effect size for comparing means from two groups was defined as:

$$\textit{estimated effect size} = \frac{|\hat{\mu}_i - \hat{\mu}_j|}{\sqrt{\dfrac{S_i^2 + S_j^2}{2}}}$$

where $\hat{\mu}_i$ refers to the estimated mean for group *i*, and S_i refers to the estimated standard deviation within group *i*. The within-group estimated standard deviations were taken to be the standard deviation of the set of five plausible values for the students in subgroup *i* and were calculated using the Westat sampling weights.

The estimated effect size for comparing proportions was defined as $|f_i - f_j|$, where $f_i = 2 \arcsin\sqrt{p_i}$, and p_i is the estimated proportion in group *i*.

For both means and proportions, no qualifying language was used in describing significant group differences when the estimated effect size exceeded .1. However, when a significant difference was found but the estimated effect size was less than .1, the qualifier *somewhat* was used. For example, if the mean proficiency for females was significantly higher than that for males but the estimated effect size of the difference was less than .1, females were described as performing *somewhat higher* than males.

The principal audience for the state reports was taken to consist of curriculum- and policy-oriented education specialists. Although it was assumed that such an audience would have some degree of familiarity with statistics, an attempt was made to keep the amount of statistical jargon to a minimum. This caused a certain degree of difficulty for group comparisons in which no statistically significant difference was obtained. In such cases, the rigorous statistical interpretation is not that the groups are the same (one does not prove the null hypothesis), but that the data are not sufficiently strong to justify concluding that a difference exists. In order to minimize the use of phrases such as "no statistically significant difference," the performance levels of the groups being compared were sometimes described as being "about the same". Readers were cautioned in the introduction to the state reports to interpret such statements to mean "no statistically significant difference."

The reliance on significance tests for commenting on differences while adopting a convention of describing null results as "about the same" resulted in situations that might appear somewhat anomalous to a reader of the report. Due to variations in subgroup sample sizes and standard errors, group differences between point estimates of one quantity (like a subgroup mean or proportions) could be large in an absolute sense but not statistically different (and hence described as "about the same"), while a considerably smaller difference between another pair of groups was described as indicating different levels of performance. An attempt was made to minimize potential confusion by footnoting large but nonsignificant differences. If the difference in proficiency means between two groups was greater than 7 points, a footnote appeared on the page on which the comparison was described. The footnote read, "Recall that 'about the same' means that the difference between groups, although it may appear large, is not statistically significant."

10.5.5 Descriptions of the Magnitude of Percentage

Percentages reported in the text of the state reports are sometimes described using quantitative words or phrases. For example, the number of students being taught by teachers with master's degrees in mathematics might be described as "relatively few" or "almost all," depending on the size of the percentage in question. Any convention for choosing descriptive terms for the magnitude of percentages is to some degree arbitrary. The rules used to select the descriptive phrases in the report are given in Table 10-2.

Table 10-2
Rules for Selecting Descriptions of Percentages

Percentage	Descriptive Text Used in Report
$p = 0$	None
$0 < p \leq 10$	Relatively few
$10 < p \leq 20$	Some
$20 < p \leq 30$	About one-quarter
$30 < p \leq 44$	Less than half
$44 < p \leq 55$	About half
$55 < p \leq 69$	More than half
$69 < p \leq 79$	About three-quarters
$79 < p \leq 89$	Many
$89 < p < 100$	Almost all
$p=100$	All

APPENDIX A

PARTICIPANTS IN THE OBJECTIVES AND ITEM DEVELOPMENT PROCESS

APPENDIX A

PARTICIPANTS IN THE OBJECTIVES AND ITEM DEVELOPMENT PROCESS

PROJECT STEERING COMMITTEE

American Association of School Administrators
Gary Marx, Associate Executive Director
Arlington, Virginia

American Educational Research Association
Carole Perlman, Director of Research and Evaluation
Chicago, Illinois

American Federation of Teachers
Marilyn Rauth, Director
Educational Issues
Washington, D.C.

Association of State Assessment Programs
Edward Roeber, Co-Chairman
Lansing, Michigan

Association of Supervision and Curriculum Development
Helene Hodges
Alexandria, Virginia

Council of Chief State School Officers
H. Dean Evans, Superintendent of Public Instruction
State Department of Education
Indianapolis, Indiana

National Alliance of Business
Esther Schaeffer
Washington, D.C.

National Association of Elementary School Principals
Kathleen Holliday, Principal
Potomac, Maryland

National Educational Association
Ann Smith, NEA Board Member
Ormond Beach, Florida

National Governors' Association
Mike Cohen
Washington, D.C.

National Parent Teacher Association
Ann Kahn
Alexandria, Virginia

National Education of Secondary School Principals
Scott Thompson, Executive Director
Reston, Virginia

National School Board Association
Harriet C. Jelnek, Director
Rhineland, Wisconsin

National Association of Test Directors
Paul Le Mahieu
Pittsburgh, Pensylvania

National Catholic Educational Association
Brother Robert Kealey
Washington, D.C.

PROJECT PLANNING COMMITTEE

Marilyn Adams
BBN Laboratories
Cambridge, Massachusetts

Marsha Delain
South Carolina
Department of Education
Columbia, South Carolina

Lisa Delpit
Institute for Urban Research
Morgan State University
Baltimore, Maryland

William Feehan
Chase Manhattan Bank
New York, New York

Philip Gough
Department of Psychology
University of Texas at Austin
Austin, Texas

Edward Haertel
Stanford University
Stanford, California

Elfrieda Hiebert
School of Education
University of Colorado
Boulder, Colorado

Judith Langer
School of Education
State University of New York, Albany
Albany, New York

P. David Pearson
University of Illinois
College of Education
Champaign, Illinois

Charles Peters
Oakland Schools
Pontiac, Michigan

John P. Pikulski
College of Education
University of Delaware
Newark, Delaware

Keith Stanovich
Oakland University
Rochester, Michigan

Paul Randy Walker
Maine Department of Education
Augusta, Maine

Sheila Valencia
University of Washington
Seattle, Washington

Janet Jones
Charles County Public Schools
Waldorf, Maryland

1992 NAEP READING CCSSO PROJECT STAFF

Ramsay W. Selden, Director
State Education Assessment Center
Council of Chief State
School Officers

Barbara Kapinus
Project Coordinator

Diane Schilder
Project Associate

THE READING ITEM DEVELOPMENT COMMITTEE

Dr. Mary Barr
San Diego, CA

Dr. Carita Chapman
Chicago, IL

Dr. Richard Halle
Marshfield, WI

Dr. Elfrieda Hebert
School of Education
University of Colorado
Boulder, CO

Dr. Judith Langer
School of Education
SUNY - Albany
Albany, NY

Edye Norniella
Miami, FL

Dr. Charles Peters
Oakland Schools
Waterford, MI

Dr. John Pikulski
Newark, DE

Dr. Robert Swartz
Newtonville, MA

Dr. Barbara Kapinus
Hyattsville, MD

APPENDIX B

SUMMARY OF PARTICIPATION RATES

Guidelines for Sample Participation and Explanation of Derivation of Weighted Participation for the 1992 Trial State Assessment in Reading

Introduction

Since 1989, state representatives, the National Assessment Governing Board (NAGB), several committees of external advisors to the National Assessment of Educational Progress (NAEP), and the National Center for Education Statistics (NCES) have engaged in numerous discussions about the procedures for reporting the NAEP Trial State Assessment results. As part of these discussions, it was recognized that sample participation rates across the states and territories have to be uniformly high to permit fair and valid comparisons. Therefore, NCES established four guidelines for school and student participation in the 1990 Trial State Assessment Program.

The participation rate data were first presented in the appendix of the 1990 composite mathematics report (*The State of Mathematics Achievement*) and a notation was made in those appendix tables and in Table 2 of the appropriate state report for any jurisdiction with participation levels that did not meet the guidelines. Virtually every state and territory met or exceeded the four guidelines for the 1990 program.

For the 1992 Trial State Assessment, NCES has decided to continue to use those four guidelines, the first two relating to school participation and the second two relating to student participation. The guidelines are based on the standards for sample surveys that are set forth in the U.S. Department of Education's *Standards and Policies* (1987). Three of the guidelines for the 1992 program are identical to those used in 1990, while one guideline for school participation has been modified.

NCES and NAGB have reviewed the policy of how participation rates can best be presented so that readers of reports can accurately assess the quality of the data being reported. They have decided that for reporting the results from the 1992 Trial State Assessment Program, tables again will have notations for the jurisdictions not meeting each guideline. They also have decided that there will be a fuller discussion in the body of the 1992 composite reports about the participation rates and nature of the samples for each of the participating jurisdictions.

The participation rate information for the 1992 Trial State Assessment of mathematics at grades 4 and 8 was presented in the document *School & Student Participation Rates for the Mathematics Assessment and Guidelines for Sample Participation*, which was distributed for the states' review on September 1992. It also will appear in appendices in both the *NAEP 1992 Mathematics Report Card for the Nation and the States* and the *Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics.* This document provides similar participation rate information for the 1992 Trial State Assessment of reading at grade 4.

The next section or this report provides an explanation of the guidelines and notations. In brief, the guidelines cover levels of school and student participation, both overall and for particular population classes. Consistent with the NCES standards, weighted data must be used to calculate all participation rates for sample surveys, and weighted rates will be provided in the reports. The procedures used to derive the weighted school and student participation rates are provided immediately after the discussion of the guidelines and notations.

The final section of this report consists of a set of tables that provide the 1992 participation rate information for the 1992 Trial State reading assessment. Because the aggregate across all states is not representative of any meaningful sample, the weighted participation rates across states have not been analyzed. However, the national and regional counts from the national assessment have been included and do provide some context for interpreting the summary of activities in each individual state and territory.

Notations for Use in Reporting School and Student Participation Rates

Unless the overall participation rate is sufficiently high for a state or territory, there is a risk that the assessment results for that jurisdiction are subject to appreciable nonresponse bias. Moreover, even if the overall participation rate is high, there may be significant nonresponse bias if the nonparticipation that does occur is heavily concentrated among certain classes of schools or students.

The following notations concerning school and student participation rates in the Trial State Assessment Program were established to address four significant ways in which nonresponse bias could be introduced into the jurisdiction sample estimates. The four conditions that will result in a state or territory receiving a notation in the 1992 reports are presented below. Note that in order to receive no notations, a state or territory must satisfy all four guidelines.

A jurisdiction will receive a notation if:

1. **Both the state's weighted participation rate for the initial sample of schools was below 85 percent AND the weighted school participation rate after substitution was below 90 percent; OR the weighted school participation rate of the initial sample of schools was below 70 percent (regardless of the participation rate after substitution.)**

Discussion: For states or territories that did not use substitute schools, the participation rates are based on participating schools from the original sample. In these situations, the NCES standards specify weighted school participation rates of at least 85 percent to guard against potential bias due to school nonresponse. Thus, the first part of the notation that refers to the weighted school participation rate for the initial sample of schools is in direct accordance with NCES standards.

To help ensure adequate sample representation for each jurisdiction participating in the 1992 Trial State Assessment Program, NAEP provided substitutes for nonparticipating schools. When possible, a substitute school was provided for each initially selected school that declined participation before November 15, 1991. For states or territories that used substitute schools, the assessment results will be based on the student data from all participating schools from both the original sample and the list of substitutes (unless both an initial school and its substitute eventually participated, in which case only the data from the initial school will be used).

The NCES standards do not explicitly address the use of substitute schools to replace initially selected schools that decide not to participate in the assessment. However, considerable technical consideration was given to this issue. Even though the characteristics of the substitute schools were matched as closely as possible to the characteristics of the initially selected schools, substitution does not entirely eliminate bias due to the nonparticipation of initially selected schools. Thus, for the weighted school participation rates including substitute schools, the guideline was set at 90 percent.

Finally, if the jurisdiction's school participation rate for the initial sample of schools is below 70 percent, even if the rate after substitution exceeds 90 percent, there is a substantial possibility that, in aggregate, the substitute schools are not sufficiently similar to the schools that they replaced to assure that there is negligible bias in the assessment results. The last part of the notation takes this into consideration.

A jurisdiction will receive a notation if:

2. **The nonparticipating schools included a class of schools with similar characteristics, which together accounted for more than five percent of the state's total fourth-grade weighted sample of public schools. The classes of schools from each of which a state needed minimum school participation levels were determined by degree of urbanization, minority enrollment, and median household income of the area in which the school is located.**

Discussion: The NCES standards specify that attention should be given to the representativeness of the sample coverage. Thus, if some important segment of the jurisdiction's population is not adequately represented, it is of concern, regardless of the overall participation rate.

This notation addresses the fact that, if nonparticipating schools are concentrated within a particular class of schools, the potential for substantial bias remains, even if the overall level of school participation appears to be satisfactory. Nonresponse adjustment cells have been formed within each jurisdiction, and the schools within each cell are similar with respect to minority enrollment, degree of urbanization, and/or median household income, as appropriate for each jurisdiction.

If more than five percent (weighted) of the sampled schools (after substitution) are nonparticipants from a single adjustment cell, then the potential for nonresponse bias is too great. This guideline is based on the NCES standard for stratum-specific school nonresponse rates.

A jurisdiction will receive a notation if:

3. **The weighted student response rate within participating schools was below 85 percent.**

Discussion: This guideline follows the NCES standard of 85 percent for overall student participation rates. The weighted student participation rate is based on all eligible students from initially selected or substitute schools who participated in the assessment in either an initial session or a make-up session. If the rate falls below 85 percent, then the potential for bias due to students' nonresponse is too great.

A jurisdiction will receive a notation if:

4. **The nonresponding students within participating schools included a class of students with similar characteristics, who together comprised more than five percent of the state's weighted assessable student sample. Student groups from which a state needed minimum levels of participation were determined by age of student and type of assessment session (unmonitored or monitored), as well as school level of urbanization, minority enrollment, and median household income of the area in which the school is located.**

Discussion: This notation addresses the fact that if nonparticipating students are concentrated within a particular class of students, the potential for substantial bias remains, even if the overall student participation level appears to be satisfactory. Student nonresponse adjustment cells have been formed using the school-level nonresponse adjustment cells, together with the student's age and the nature of the assessment session (unmonitored or monitored). If more than five percent (weighted) of the invited students who do not participate in the assessment are from a single adjustment cell, then the potential for nonresponse bias is too great. This guideline is based on the NCES standard for stratum-specific student nonresponse rates.

Derivation of Weighted Participation Rates

Weighted School Participation Rates. The weighted school participation rates within each state or territory provide the percentages of fourth-grade students in public schools who are represented by the schools participating in the assessment, prior to statistical adjustments for school nonresponse.

Two weighted school participation rates are computed for each state and territory. The first is the weighted participation rate for the initial sample of schools. This rate is based only on those schools that were initially selected for the assessment. The numerator of this rate is the sum of the number of students represented by each initially selected school that participated in the assessment. The denominator is the sum of the number of students represented by each of the initially selected schools found to have eligible students enrolled. This includes both participating and nonparticipating schools.

The second participation rate is the weighted participation rate after substitution. The numerator of this rate is the sum of the number of students represented by each of the participating schools, whether originally selected or a substitute. The denominator is the same as that for the weighted participation rate for the initial sample. This means that, for a given state, grade, and subject, the weighted participation rate after substitution is always at least as great as the weighted participation rate for the initial sample of schools.

In general, different schools in the sample can represent different numbers of students in the state population. The number of students represented by an initially selected school (the school weight) is the fourth-grade enrollment of the school divided by the probability that the school was included in the sample. For instance, a selected school with a fourth-grade enrollment of 150 and a selection probability of 0.2 represents 750 students from that state. The number of students represented by a substitute school is the number of students represented by the replaced nonparticipating school.

Because each selected school represents different numbers of students in the population, the weighted school participation rates may differ somewhat from the simple unweighted rates. (The unweighted rates are calculated from the counts of school by dividing the number of participating schools by the number of schools in the sample.) The difference between the weighted and the unweighted rates is potentially largest in smaller jurisdictions where all schools with fourth-grade students were included in the sample. In those jurisdictions, each school represents only its own students. Therefore, the nonparticipation of a large school reduces the weighted school participation rate by a greater amount than does the nonparticipation of a small school.

The nonparticipation of larger schools also has greater impact than that of smaller schools on reducing weighted school participation rates in larger jurisdictions where fewer than all of the schools were included in the sample. However, since the number of students represented by each school is more nearly constant in larger states, the difference between the impact of nonparticipation by either large or small schools is less marked than in states where all schools were selected.

In general, the larger the jurisdiction, the less the difference is between the weighted and unweighted school participation rates. However, even in the smaller jurisdictions, the differences tend to be small.

Weighted Student Participation Rate. The weighted student participation rate provides the percentage of the eligible student population from participating schools within the state or territory that are represented by the students who participated in the assessment (in either an initial session or a make-up session). The eligible student population from participating schools within a jurisdiction consists of all public-school students who were in the fourth grade, who attended a school that, if selected, would have participated and who, if selected, would not have been excluded from the assessment. The numerator of this rate is the sum, across all assessed students, of the number of students represented by each assessed student (prior to adjustment for student nonparticipation). The denominator is the sum of the number of students represented by each selected student who was invited and eligible to participate (i.e., not excluded), including students who did not participate. Thus, the denominator is an estimate of the total number of assessable students in the group of schools within the jurisdiction that would have participated if selected.

The number of students represented by a single selected student (the student weight) is 1.0 divided by the overall probability that the student was selected for assessment. In general, the number of students from a jurisdiction's population that are represented by a sampled student is approximately constant across students. Consequently, there is little difference between the weighted student participation rate and the unweighted student participation rate.

Weighted Overall School and Student Participation Rate. An overall indicator of the effect of nonparticipation by both students and schools is given by the overall participation rate. This is calculated as the product of the weighted school participation rate (after substitution), and the weighted student participation rate. For jurisdictions having a high overall participation rate the potential is low for bias to be introduced through either school nonparticipation or student nonparticipation. This rate provides a summary measure that indicates the proportion of the jurisdiction's fourth-grade student population that is directly represented by the final student sample. When the overall rate is high, the adjustments for nonresponse that are used in deriving the final survey weights are likely to be effective in maintaining nonresponse bias at a negligible level. Conversely, when the overall rate is relatively low there is a greater chance that a non-negligible bias remains even after making such adjustments.

The overall rate is not used in establishing the guidelines/notations for school and student participation, since guidelines exist already covering school and student participation separately. The overall participation rate was not reported in 1990.

Derivation of Weighted Percentages for Excluded Students

Weighted Percentage of Excluded Students. The weighted percentage of excluded students estimates the percentage of the fourth-grade population in the jurisdiction's public schools that are represented by the students who were excluded from the assessment, after accounting for school nonparticipation. The numerator is the sum, across all excluded students, of the number of students represented by each excluded student. The denominator is the sum of the number of students represented by each of the students who was sampled (and had not withdrawn from the school at the time of the assessment).

Weighted Percentage of Students with an Individualized Education Plan (IEP). The weighted percentage of IEP students estimates the percentage of the fourth-grade population in the jurisdiction's public schools that are represented by the students who were classified as IEP, after accounting for school nonparticipation. The numerator is the sum, across all students classified as IEP, of the number of students represented by each IEP student. The denominator is the sum of the number of students represented by each of the students who was sampled (and had not withdrawn from the school at the time of the assessment).

Weighted Percentage of Excluded IEP Students. The weighted percentage of IEP students who were excluded estimates the percentage of students in the jurisdiction that are represented by those IEP students who were excluded from the assessment, after accounting for school nonparticipation. The numerator is the sum, across all students classified as IEP and excluded from the assessment, of the number of students represented by each excluded IEP student. The denominator is the sum of the number of students represented by each of the IEP students who was sampled (and had not withdrawn from the school at the time of the assessment).

Weighted Percentage of Limited English Proficiency (LEP) Students. The weighted percentage of LEP students estimates the percentage of the fourth-grade population in the jurisdiction's public schools that are represented by the students who were classified as LEP, after accounting for school nonparticipation. The numerator is the sum, across all students classified as LEP, of the number of students represented by each LEP student. The denominator is the sum of the number of students represented by each of the students who was sampled (and had not withdrawn from the school at the time of the assessment).

Weighted Percentage of Excluded LEP Students. The weighted percentage of LEP students who were excluded estimates the percentage of students in the jurisdiction that are represented by those LEP students who were excluded from the assessment, after accounting for school nonparticipation. The numerator is the sum, across all students classified as LEP and excluded from the assessment, of the number of students represented by each excluded LEP student. The denominator is the sum of the number of students represented by each of the LEP students who was sampled (and had not withdrawn from the school at the time of the assessment).

Note: All percentages are based on student weights that have been adjusted for school-level nonresponse.

TABLE B.4 | School Participation Rates, Grade 4, 1992 Reading Assessment

PUBLIC SCHOOLS	Weighted Percentage School Participation Before Substitution	Weighted Percentage School Participation After Substitution	Number Schools in Original Sample	Number Schools Not Eligible	Number Schools in Original Sample That Participated	Number Substituted Schools Provided	Number Substituted Schools That Participated	Total Number Schools That Participated
NATION	86	87	284	2	247	7	2	249
Northeast	80	80	56	0	46	1	0	46
Southeast	92	93	70	1	65	1	1	66
Central	92	92	64	0	59	0	0	59
West	82	83	94	1	77	5	1	78
STATES								
Alabama	76	97	112	3	82	25	23	105
Arizona[4]	99	99	107	1	106	0	0	106
Arkansas[4]	87	96	120	2	105	12	11	116
California	92	97	115	3	103	6	6	109
Colorado	100	100	124	2	122	0	0	122
Connecticut	99	99	113	4	108	0	0	108
Delaware[2 3]	92	92	56	6	44	0	0	44
Dist. Columbia	99	99	118	4	113	0	0	113
Florida	100	100	111	1	110	0	0	110
Georgia	100	100	109	2	107	0	0	107
Hawaii	100	100	106	0	106	0	0	106
Idaho	82	96	123	1	100	19	15	115
Indiana	77	92	116	2	88	24	16	104
Iowa	100	100	133	4	129	0	0	129
Kentucky[4]	94	97	124	3	116	3	3	119
Louisiana	100	100	115	4	111	0	0	111
Maine[1 2 4 5]	58	71	141	1	76	41	20	96
Maryland	99	99	112	1	110	1	0	110
Massachusetts	87	97	123	4	103	12	11	114
Michigan[4]	83	90	116	3	92	17	8	100
Minnesota[5]	81	94	116	5	91	15	13	104
Mississippi	98	100	110	3	105	2	2	107
Missouri	90	97	123	6	105	9	9	114
Nebraska[1 2]	76	87	161	7	106	41	15	121
New Hampshire[1 2 4 5]	68	81	128	4	83	34	17	100
New Jersey[1 2]	76	82	121	4	89	23	7	96
New Mexico[4 5]	76	91	114	1	84	26	18	102
New York[1 2 4]	78	84	110	0	86	21	7	93
North Carolina[4]	95	99	118	2	111	5	5	116
North Dakota	70	91	133	3	97	33	23	120
Ohio	78	91	121	1	93	21	15	108
Oklahoma	86	98	130	0	115	14	13	128
Pennsylvania	85	95	119	0	102	17	12	114
Rhode Island	83	96	114	5	89	15	15	104
South Carolina	98	99	112	1	109	1	1	110
Tennessee	93	94	120	1	110	8	1	111
Texas	92	97	111	3	98	5	5	103
Utah	99	99	110	1	108	0	0	108
Virginia	99	99	118	4	113	0	0	113
West Virginia	100	100	144	7	137	0	0	137
Wisconsin[4]	99	99	127	5	122	0	0	122
Wyoming	97	97	158	6	148	0	0	148
TERRITORY								
Guam[3]	100	100	21	0	21	0	0	21

See explanations of the notations and guidelines about sample representativeness and for the derivation of weighted participation. [1] Both the state's weighted participation rate for the initial sample of schools was below 85% AND the weighted school participation rate after substitution was below 90%; OR the weighted school participation rate of the initial sample of schools was below 70% (regardless of the participation rate after substitution.) [2] The nonparticipating schools included a class of schools with similar characteristics, which together accounted for more than five percent of the state's total fourth- or eighth-grade weighted sample of public schools. The classes of schools from each of which a state needed minimum school participation levels were determined by urbanicity, minority enrollment, and median household income of the area in which the school is located. [3] The Trial State Assessment was based on all eligible schools. There was no sampling of schools. [4] In one or more schools an assessment was conducted, but either the wrong materials were sent to the school(s) or the materials were lost in shipping via the U.S. Postal Service. The school(s) are included in the counts of participating schools, both before and after substitution. However, in the weighted results, the school(s) are treated in the same manner as a nonparticipating school because no student responses were available for analysis and reporting. [5] One or more schools in the original sample initially declined and then decided to participate after their substitute(s) had also agreed to participate. Further, assessments were conducted in both the original and substitute schools. For these cases the substitute school is included in the number of substitute schools provided and in the number of substitute schools participating. The state's estimates will be based on the student responses from the original school only.

SOURCE: National Assessment of Educational Progress (NAEP), 1992 Reading Assessment.

TABLE B.5 | Student Participation Rates, Grade 4, 1992 Reading Assessment

PUBLIC SCHOOLS	Weighted Percentage Student Participation After Make-ups	Number Students Original Sample	Number Students Supplemental Sample	Number Students Withdrawn	Number Students Excluded	Number Students to be Assessed	Number Students Assessed Initial Sessions	Number Students Assessed Make-ups	Total Number Students Assessed
NATION	94	5,981	--	--	602	5,379	5,038	7	5,045
Northeast	95	1,055	--	--	104	951	903	0	903
Southeast	94	1,595	--	--	128	1,467	1,381	1	1,382
Central	95	1,281	--	--	71	1,210	1,137	6	1,143
West	93	2,050	--	--	299	1,751	1,617	0	1,617
STATES									
Alabama	96	2,885	58	106	153	2,684	2,567	4	2,571
Arizona[2]	95	3,095	146	216	218	2,807	2,659	18	2,677
Arkansas[2]	96	2,909	87	144	153	2,699	2,585	4	2,589
California	94	3,041	139	234	440	2,506	2,345	20	2,365
Colorado	95	3,275	129	160	204	3,040	2,882	15	2,897
Connecticut	95	2,914	52	106	205	2,655	2,506	8	2,514
Delaware	95	2,330	90	126	138	2,156	2,040	8	2,048
Dist. Columbia	94	3,033	76	177	284	2,648	2,472	24	2,496
Florida	95	3,258	187	224	296	2,925	2,751	16	2,767
Georgia	96	3,078	115	202	159	2,832	2,705	7	2,712
Hawaii	95	2,995	121	154	171	2,791	2,624	18	2,642
Idaho	96	2,934	88	121	112	2,789	2,671	3	2,674
Indiana	96	2,798	69	103	114	2,650	2,532	3	2,535
Iowa	96	3,006	49	80	115	2,860	2,747	9	2,756
Kentucky	96	3,007	111	143	112	2,863	2,728	24	2,752
Louisiana	96	3,159	98	145	135	2,977	2,834	14	2,848
Maine[1]	95	2,183	27	49	123	2,038	1,932	7	1,939
Maryland	95	3,193	123	199	199	2,918	2,762	4	2,786
Massachusetts	96	2,935	29	77	224	2,663	2,535	10	2,545
Michigan[2]	94	2,777	71	97	136	2,615	2,436	10	2,446
Minnesota[1]	96	2,895	35	72	117	2,741	2,607	13	2,620
Mississippi	97	2,981	99	177	150	2,753	2,649	8	2,657
Missouri	95	2,834	129	153	124	2,686	2,548	14	2,562
Nebraska	96	2,648	46	72	126	2,496	2,383	10	2,393
New Hampshire	96	2,554	53	75	115	2,417	2,314	8	2,322
New Jersey	96	2,510	62	91	139	2,342	2,221	18	2,239
New Mexico[1]	95	2.852	71	201	214	2,508	2,380	2	2,382
New York	95	2,594	49	76	149	2,418	2,278	7	2,285
North Carolina	96	3,128	129	130	136	2,991	2,871	12	2,883
North Dakota	97	2,275	34	39	48	2,222	2,158	0	2,158
Ohio	96	2,910	90	117	179	2,704	2,580	0	2,580
Oklahoma	85	2,936	115	153	240	2,658	2,251	0	2,251
Pennsylvania	95	3,071	69	77	122	2,941	2,791	14	2,805
Rhode Island	95	2,764	58	166	192	2,464	2,344	3	2,347
South Carolina	96	3,083	116	172	170	2,857	2,758	0	2,758
Tennessee	95	3,047	127	159	141	2,874	2,728	6	2,734
Texas	96	2,987	106	163	252	2,678	2,567	4	2,571
Utah	96	3,139	94	159	140	2,934	2,819	10	2,829
Virginia	96	3,128	117	132	199	2,914	2,782	4	2,786
West Virginia	96	3,009	80	89	152	2,848	2,722	11	2,733
Wisconsin[2]	96	3,049	49	72	199	2,827	2,712	0	2,712
Wyoming	96	3,046	124	152	124	2,894	2,775	0	2,775
TERRITORY									
Guam	94	2,268	134	94	154	2,154	2,025	4	2,029

See explanations of the notations and guidelines about sample representativeness and for the derivation of weighted participation. [1] One or more schools in the original sample initially declined and then decided to participate after their substitute(s) had also agreed to participate. Further, assessments were conducted in both the original and substitute schools. For these cases, the students in the substitute school(s) are included in the counts of students in the table. The state's estimates will be based on the student responses from the original school only. [2] In one or more schools an assessment was conducted but the wrong materials were sent to the school(s). The students in these school(s) are included in the counts of students in the tables. However, the state's estimates will not be based on these student responses. (--) Because student sampling for the national assessment was implemented within several days of the assessment within each school there was no supplemental sample and the number of students withdrawn was negligible.

SOURCE: National Assessment of Educational Progress (NAEP), 1992 Reading Assessment.

TABLE B.6 | Summary of School and Student Participation, Grade 4, 1992 Reading Assessment

PUBLIC SCHOOLS	Weighted Percentage School Participation Before Substitution	Weighted Percentage School Participation After Substitution	Notation Number 1	Weighted Percentage Student Participation After Make-ups	Notation Number 3	Weighted Overall Rate
NATION	86	87		94		82
Northeast	80	80		95		76
Southeast	92	93		94		87
Central	92	92		95		87
West	82	83		93		77
STATES						
Alabama	76	97		96		93
Arizona	99	99		95		95
Arkansas	87	96		96		93
California	92	97		94		92
Colorado	100	100		95		95
Connecticut	99	99		95		94
Delaware*	92	92		95		88
Dist. Columbia	99	99		94		94
Florida	100	100		95		95
Georgia	100	100		96		96
Hawaii	100	100		95		95
Idaho	82	96		96		92
Indiana	77	92		96		88
Iowa	100	100		96		96
Kentucky	94	97		96		93
Louisiana	100	100		96		96
Maine*	58	71	***	95		67
Maryland	99	99		95		95
Massachusetts	87	97		96		92
Michigan	83	90		94		84
Minnesota	81	94		96		90
Mississippi	98	100		97		97
Missouri	90	97		95		93
Nebraska*	76	87	***	96		83
New Hampshire*	68	81	***	96		77
New Jersey*	76	82	***	96		79
New Mexico	76	91		95		86
New York*	78	84	***	95		79
North Carolina	95	99		96		95
North Dakota	70	91		97		89
Ohio	78	91		96		87
Oklahoma	86	98		85		83
Pennsylvania	85	95		95		91
Rhode Island	83	96		95		92
South Carolina	98	99		96		96
Tennessee	93	94		95		89
Texas	92	97		96		93
Utah	99	99		96		95
Virginia	99	99		96		95
West Virginia	100	100		96		96
Wisconsin	99	99		96		95
Wyoming	97	97		96		93
TERRITORY						
Guam	100	100		94		94

See explanations of the notations and guidelines about sample representativeness and for the derivation of weighted participation. **Notation Number 1** = Both the state's weighted participation rate for the initial sample of schools was below 85% AND the weighted school participation rate after substitution was below 90%; OR the weighted school participation rate of the initial sample of schools was below 70% (regardless of the participation rate after substitution.) **Notation number 3** = The weighted student response rate within participating schools was below 85 percent.

SOURCE: National Assessment of Educational Progress (NAEP), 1992 Reading Assessment.

TABLE B.7 | Weighted Percentages of Students Excluded (IEP and LEP) from Original Sample, Grade 4, 1992 Reading Assessment

PUBLIC SCHOOLS	Total Percentage Students Identified IEP and LEP	Total Percentage Students Excluded	Percentage Students Identified IEP	Percentage Students Excluded IEP	Percentage Students Identified LEP	Percentage Students Excluded LEP
NATION	12	8	9	6	4	3
Northeast	12	8	9	5	3	3
Southeast	11	7	9	6	1	1
Central	7	5	6	4	1	1
West	18	12	10	6	9	7
STATES						
Alabama	10	6	10	5	0	0
Arizona	16	7	8	5	10	3
Arkansas	11	5	11	5	0	0
California	28	14	7	4	21	11
Colorado	11	6	9	5	2	2
Connecticut	15	7	12	4	4	3
Delaware*	12	6	11	5	1	0
Dist. Columbia	12	10	9	7	4	3
Florida	17	9	14	7	4	2
Georgia	9	5	8	5	1	1
Hawaii	14	6	9	4	5	2
Idaho	9	4	8	3	2	1
Indiana	8	4	7	4	0	0
Iowa	10	4	9	4	1	0
Kentucky	8	4	7	4	0	0
Louisiana	8	4	7	4	1	0
Maine*	12	5	12	5	0	0
Maryland	14	7	12	6	2	1
Massachusetts	17	7	14	5	4	2
Michigan	7	5	6	4	1	1
Minnesota	10	4	8	4	2	1
Mississippi	7	5	7	5	0	0
Missouri	11	5	11	4	0	0
Nebraska*	13	4	13	4	1	1
New Hampshire*	12	4	12	4	0	0
New Jersey*	10	6	7	3	4	2
New Mexico	14	8	10	6	3	2
New York*	13	6	8	4	5	2
North Carolina	12	4	11	4	1	1
North Dakota	10	2	10	2	0	0
Ohio	10	6	9	6	1	1
Oklahoma	13	8	12	8	2	1
Pennsylvania	9	4	8	3	2	1
Rhode Island	16	7	10	4	6	4
South Carolina	11	6	11	6	0	0
Tennessee	12	5	11	5	0	0
Texas	17	8	9	5	9	3
Utah	10	4	9	4	1	1
Virginia	12	6	11	6	1	1
West Virginia	8	5	8	5	0	0
Wisconsin	11	7	9	6	2	1
Wyoming	11	4	10	4	1	0
TERRITORY						
Guam	12	7	6	4	6	3

IEP = Individual Education Plan and LEP = Limited English Proficiency. To be excluded, a student was supposed to be IEP or LEP and judged incapable of participating in the assessment. A student reported as both IEP and LEP is counted once in the overall rate (first column), once in the overall excluded rate (second column), and separately in the remaining columns. Note: Weighted percentages for the nation and region are based on students sampled for all subject areas assessed in 1992 (mathematics, reading, and writing). However, based on the national sampling design, the rates shown also are the best estimates for the reading assessment.

SOURCE: National Assessment of Educational Progress (NAEP), 1992 Reading Assessment.

BEST COPY AVAILABLE

TABLE B.8 | Weighted Percentages of Absent, IEP, and LEP Students Based on Those Invited to Participate in the Assessment, Grade 4, 1992 Reading Assessment

PUBLIC SCHOOLS	Weighted Percentage Student Participation After Make-up	Weighted Percentage Absent	Weighted Percentage Assessed IEP	Weighted Percentage Absent IEP	Weighted Percentage Assessed LEP	Weighted Percentage Absent LEP
NATION	94	6	89	11	93	7
Northeast	94	6	93	7	81	19
Southeast	93	7	83	17	68	32
Central	94	6	92	8	96	4
West	93	7	90	10	94	6
STATES						
Alabama	96	4	92	8	68	32
Arizona	95	5	93	7	95	5
Arkansas	96	4	94	6	100	0
California	94	6	95	5	94	6
Colorado	95	5	89	11	90	10
Connecticut	95	5	91	9	94	6
Delaware*	95	5	95	5	100	0
Dist. Columbia	94	6	92	8	93	7
Florida	95	5	90	10	96	4
Georgia	96	4	89	11	100	0
Hawaii	95	5	89	11	98	2
Idaho	96	4	91	9	95	5
Indiana	96	4	93	7	100	0
Iowa	96	4	95	5	100	0
Kentucky	96	4	95	5	100	0
Louisiana	96	4	92	8	100	0
Maine*	95	5	93	7	80	20
Maryland	95	5	94	6	94	6
Massachusetts	96	4	93	7	97	3
Michigan	94	6	80	20	92	8
Minnesota	96	4	93	7	100	0
Mississippi	97	3	93	7	100	0
Missouri	95	5	94	6	100	0
Nebraska*	96	4	95	5	88	12
New Hampshire*	96	4	92	8	78	22
New Jersey*	96	4	97	3	97	3
New Mexico	95	5	84	16	93	7
New York*	95	6	96	4	98	2
North Carolina	96	4	94	6	89	11
North Dakota	97	3	97	3	100	0
Ohio	96	4	91	9	100	0
Oklahoma	85	15	73	27	88	12
Pennsylvania	95	4	93	7	94	6
Rhode Island	95	5	97	3	97	3
South Carolina	96	4	93	7	0	0
Tennessee	95	5	93	7	69	31
Texas	96	4	95	5	97	3
Utah	96	4	98	2	86	14
Virginia	96	4	94	6	95	5
West Virginia	96	4	97	3	100	0
Wisconsin	96	4	95	5	100	0
Wyoming	96	4	94	6	100	0
TERRITORY						
Guam	94	6	84	16	98	2

IEP = Individual Education Plan and LEP = Limited English Proficiency. Note: Weighted percentages for the nation and region are based on students sampled for all subject areas assessed in 1992 (mathematics, reading, and writing). However, based on the national sampling design, the rates shown also are the best estimates for the reading assessment.

SOURCE: National Assessment of Educational Progress (NAEP), 1992 Reading Assessment.

TABLE B.9 | Questionnaire Response Rates, Grade 4, 1992 Reading Assessment

PUBLIC SCHOOLS	Weighted Percentage of Students Matched to Reading Teacher Questionnaires	Percentage of Reading Teacher Questionnaires Returned	Weighted Percentage of Students Matched to School Characteristics / Policies Questionnaire	Percentage of School Characteristics / Policies Questionnaires Returned	Percentage of Excluded Student Questionnaires Returned
NATION	72.3	97.7	98.9	98.4	91.0
Northeast	75.6	95.8	100.0	100.0	94.6
Southeast	80.4	99.0	95.6	95.5	94.4
Central	74.9	97.6	99.7	98.3	93.3
West	60.4	97.2	100.0	100.0	87.1
STATES					
Alabama	90.4	100.0	100.0	100.0	100.0
Arizona	90.1	99.6	99.0	99.1	97.7
Arkansas	93.2	100.0	99.3	99.1	98.7
California	88.9	99.3	98.7	99.1	90.7
Colorado	82.2	99.3	100.0	100.0	97.1
Connecticut	87.5	99.8	98.5	98.1	83.9
Delaware*	95.5	100.0	100.0	100.0	99.3
Dist. Columbia	73.8	99.0	93.7	94.7	90.8
Florida	88.4	98.9	99.3	99.1	97.3
Georgia	86.9	99.3	100.0	100.0	96.9
Hawaii	92.2	98.8	98.8	99.1	97.1
Idaho	93.3	99.7	100.0	100.0	100.0
Indiana	89.2	100.0	100.0	100.0	99.1
Iowa	89.8	99.5	100.0	100.0	98.3
Kentucky	90.0	99.5	99.4	99.1	100.0
Louisiana	87.5	99.6	98.2	98.2	98.5
Maine*	85.8	99.1	97.7	97.8	92.4
Maryland	90.0	99.5	100.0	100.0	95.5
Massachusetts	88.1	100.0	100.0	100.0	92.9
Michigan	87.2	100.0	100.0	100.0	97.8
Minnesota	74.7	97.6	95.7	96.1	88.9
Mississippi	86.7	99.8	100.0	100.0	99.3
Missouri	89.7	99.7	100.0	100.0	94.4
Nebraska*	82.1	100.0	99.0	99.2	98.4
New Hampshire*	93.7	99.7	97.7	99.0	99.1
New Jersey*	92.1	100.0	100.0	100.0	95.7
New Mexico	81.1	99.0	100.0	100.0	93.9
New York*	88.8	99.0	99.5	98.9	97.3
North Carolina	89.5	100.0	99.2	99.1	98.5
North Dakota	90.4	100.0	100.0	100.0	100.0
Ohio	86.9	99.5	99.7	99.1	97.2
Oklahoma	91.5	99.1	98.0	98.4	94.2
Pennsylvania	90.7	100.0	100.0	100.0	99.2
Rhode Island	88.2	99.4	99.0	98.9	95.3
South Carolina	94.2	99.6	100.0	100.0	98.8
Tennessee	91.2	100.0	98.7	98.2	95.7
Texas	85.3	99.9	99.4	99.0	99.2
Utah	91.6	99.5	100.0	100.0	100.0
Virginia	91.1	99.6	97.8	97.3	95.5
West Virginia	87.1	100.0	100.0	100.0	97.4
Wisconsin	89.5	99.7	99.5	99.2	98.5
Wyoming	85.4	100.0	100.0	100.0	99.2
TERRITORY					
Guam	92.5	98.3	93.7	95.2	99.4

The Mathematics Teacher Questionnaire requested background information about the teacher (Part I) and information about instruction in particular classes (Part II). The percentage of students matched to questionnaires is provided for Part II. If they differed, the match rates for Part I were higher. Note: For the nation and regions, the percentage of excluded student questionnaires returned is based on students sampled for all subjects assessed in 1992 (mathematics, reading, and writing). However, based on the sampling design, these rates also are the best estimates of the comparable rates for the reading assessment.

SOURCE: National Assessment of Educational Progress (NAEP), 1992 Reading Assessment.

APPENDIX C

CONDITIONING VARIABLES AND CONTRAST CODINGS

APPENDIX C

Conditioning Variables and Contrast Codings

This appendix contains information about the conditioning variables used in the construction of plausible values for the 1992 Trial State Assessment Program in reading. Two kinds of conditioning variables were defined—continuous or quasi-continuous variables, such as school mathematics score or number of hours spent watching television, and categorical variables which made up the majority of the conditioning variables created from responses to student, teacher, and school demographic and background questionnaires.

Categorical conditioning variables derived from questionnaire or demographic variables were incorporated into the conditioning process by constructing a set of contrasts, each of which defines one or more of the variable's response options. A recoding procedure explodes the raw student responses into a binary series of one-degree-of-freedom "dummy" variables. Questionnaire or demographic variables that possess ordinal response options, such as number of hours spent watching television, were included in the conditioning process by creating linear and/or quadratic multi-degree-of-freedom contrasts. Continuous variables were included in the conditioning process in their original form.

The remainder of this appendix gives the specifications used for constructing the conditioning variables. Table C-1 defines the information provided for each variable.

As described in Chapter 9, the linear conditioning model employed for the estimation of plausible values in each jurisdiction did not directly use the conditioning variable specifications listed in this appendix. To eliminate inherent instabilities in estimation encountered when using a large number of correlated variables, a principal component transformation of the correlation matrix obtained from the conditioning variable contrasts derived according to these primary specifications was performed. The principal components scores based on this transformation were used as the predictor variables in estimating the linear conditioning model.

Table C-1
Description of Data Provided for Each Conditioning Variable

Title	Description
CONDITIONING ID	An unique eight-character ID assigned to identify each conditioning variable corresponding to a particular background or subject area question within the entire pool of conditioning variables. The first four characters identify the origin of the variable: BACK (background questionnaire), READ (student reading questionnaire), SCHL (school questionnaire), TCHR (background part of teacher questionnaire), and TRED (reading classroom part of teacher questionnaire). The second four digits represent the sequential position within each origin group.
DESCRIPTION	A short description of the conditioning variable.
GRADES/ASSESSMENTS	Three characters identifying assessment ("S" for state, "N" for national) and grade (04, 08, and 12) in which the conditioning variable was used.
GROUP LABEL	A descriptive eight-character label identifying the conditioning variable.
NAEP ID	The seven-character NAEP database identification for the conditioning variable.
TYPE OF CONTRAST	The type of conditioning variable. "CLASS" identifies a categorical conditioning variable and "SCALE" identifies continuous or quasi-continuous conditioning variables.
LENGTH OF CONTRAST FIELD	The number of columns (or length of the contrast field) for the conditioning variable within the entire conditioning variable vector. The length is associated with the number of explicit contrasts comprising categorical conditioning variables.
DEGREES OF FREEDOM	The number of degrees of freedom for each contrast constructed for the conditioning variable.
NUMBER OF SPECIFICATION RECORDS	The number of unique contrasts corresponding to each conditioning variable. For each contrast a specifications record is given with the following information: a sequential identification number, an eight-character descriptive label corresponding to the associated questionnaire option(s), a "collapsing code string" enclosed in parentheses specifying the database values to be merged to form the contrast, the contrast itself, and a short description of the contrast.

```
CONDITIONING ID:          BACK0001
DESCRIPTION:              GRAND MEAN
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              OVERALL             LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                  BKSER               DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         SCALE               NUMBER OF SPECIFICATION RECORDS:  1

001  OVERALL    (@          )  1              GRAND MEAN

CONDITIONING ID:          BACK0002
DESCRIPTION:              GENDER (DERIVED)
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              GENDER              LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                  DSEX                DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  2

001  MALE       (1          )  0              GENDER:  MALE
002  FEMALE     (2          )  1              GENDER:  FEMALE

CONDITIONING ID:          BACK0003
DESCRIPTION:              ETHNICITY/RACE (DERIVED)
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              ETHNICTY            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                  DRACE               DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  4

001  WHIT/AOM   (1,5,6,M    )  000            ETHNICITY:  WHITE, AMERICAN INDIAN, UNCLASSIFIED, MISSING
002  BLACK      (2          )  100            ETHNICITY:  BLACK
003  HISPANIC   (3          )  010            ETHNICITY:  HISPANIC
004  ASIAN      (4          )  001            ETHNICITY:  ASIAN AMERICAN

CONDITIONING ID:          BACK0005
DESCRIPTION:              TYPE OF COMMUNITY (STATE ONLY)
GRADES/ASSESSMENTS:       S04, S08
GROUP LABEL:              TOC                 LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                  TOC                 DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  3

001  TOC-OTHR   (1,4,M      )  00             TOC:  EXTREME RURAL, OTHER, MISSING
002  LO_METRO   (2          )  10             TOC:  LOW METROPOLITAN
003  HI_METRO   (3          )  01             TOC:  HIGH METROPOLITAN

CONDITIONING ID:          BACK0007
DESCRIPTION:              PARENTS' HIGHEST LEVEL OF EDUCATION
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              PARED               LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                  PARED               DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  5

001  <HI_SCH    (1          )  0000           PARED:  LESS THAN HIGH SCHOOL
002  HS_GRAD    (2          )  1000           PARED:  HIGH SCHOOL GRADUATE
003  POST_HS    (3          )  0100           PARED:  POST HIGH SCHOOL
004  COL_GRAD   (4          )  0010           PARED:  COLLEGE GRADUATE
005  PARED-?    (M,IDK      )  0001           PARED:  MISSING, I DON'T KNOW

CONDITIONING ID:          BACK0008
DESCRIPTION:              ITEMS IN THE HOME (NEWSPAPER, > 25 BOOKS, ENCYCLOPEDIA, MAGAZINES)
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              HOMEITMS            LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                  HOMEEN2             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  3

001  HITEM<=2   (1,M        )  00             ITEMS IN HOME:  ZERO TO TWO ITEMS, MISSING
002  HITEM=3    (2          )  10             ITEMS IN HOME:  THREE ITEMS
003  HITEM=4    (3          )  01             ITEMS IN HOME:  FOUR ITEMS

CONDITIONING ID:          BACK0009
DESCRIPTION:              HOURS OF TV WATCHING (LINEAR)
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
```

```
GROUP LABEL:              TVWATCHL        LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                  B001801         DEGREES OF FREEDOM PER CONTRAST:  6
TYPE OF CONTRAST:         SCALE           NUMBER OF SPECIFICATION RECORDS:  7

001   TV-LIN1    (1         )  0          TV WATCHING (LINEAR):  NONE
002   TV-LIN2    (2         )  1          TV WATCHING (LINEAR):  ONE HOUR OR LESS PER DAY
003   TV-LIN3    (3         )  2          TV WATCHING (LINEAR):  TWO HOURS PER DAY
004   TV-LIN4    (4,M       )  3          TV WATCHING (LINEAR):  THREE HOURS PER DAY
005   TV-LIN5    (5         )  4          TV WATCHING (LINEAR):  FOUR HOURS PER DAY
006   TV-LIN6    (6         )  5          TV WATCHING (LINEAR):  FIVE HOURS PER DAY
007   TV-LIN7    (7         )  6          TV WATCHING (LINEAR):  SIX OR MORE HOURS PER DAY

CONDITIONING ID:          BACK0010
DESCRIPTION:              HOURS OF TV WATCHING (QUADRATIC)
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              TVWATCHQ        LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                  B001801         DEGREES OF FREEDOM PER CONTRAST:  6
TYPE OF CONTRAST:         SCALE           NUMBER OF SPECIFICATION RECORDS:  7

001   TV-QUAD1   (1         )  00         TV WATCHING (QUADRATIC):  NONE
002   TV-QUAD2   (2         )  01         TV WATCHING (QUADRATIC):  ONE HOUR OR LESS PER DAY
003   TV-QUAD3   (3         )  04         TV WATCHING (QUADRATIC):  TWO HOURS PER DAY
004   TV-QUAD4   (4,M       )  09         TV WATCHING (QUADRATIC):  THREE HOURS PER DAY
005   TV-QUAD5   (5         )  16         TV WATCHING (QUADRATIC):  FOUR HOURS PER DAY
006   TV-QUAD6   (6         )  25         TV WATCHING (QUADRATIC):  FIVE HOURS PER DAY
007   TV-QUAD7   (7         )  36         TV WATCHING (QUADRATIC):  SIX OR MORE HOURS PER DAY

CONDITIONING ID:          BACK0011
DESCRIPTION:              HOME LANGUAGE MINORITY (HOW OFTEN DO PEOPLE IN HOME SPEAK  OTHER THAN ENGLISH?)
GRADES/ASSESSMENTS:       N04, S04, N08, S08, N12
GROUP LABEL:              HOMELANG        LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                  B003201         DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS           NUMBER OF SPECIFICATION RECORDS:  2

001   HL-NEV/?   (1,M       )  0          HOME LANGUAGE MINORITY:  NEVER, MISSING
002   HL-SM/AL   (2,3       )  1          HOME LANGUAGE MINORITY:  SOMTIMES, ALWAYS

CONDITIONING ID:          BACK0012
DESCRIPTION:              HOMEWORK ASSIGNED? (GRADE 4)
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              HW-CORE4        LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                  B006601         DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS           NUMBER OF SPECIFICATION RECORDS:  3

001   HW4-MISS   (M         )  00         HOMEWORK ASSIGNED?:  MISSING
002   HW4-NONE   (1         )  10         HOMEWORK ASSIGNED?:  NO HOMEWORK ASSIGNED
003   HW4-YES    (2-5       )  01         HOMEWORK ASSIGNED?:  YES

CONDITIONING ID:          BACK0013
DESCRIPTION:              AMOUNT OF HOMEWORK (LINEAR) (GRADE 4)
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              HMWRKL4         LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                  B006601         DEGREES OF FREEDOM PER CONTRAST:  3
TYPE OF CONTRAST:         SCALE           NUMBER OF SPECIFICATION RECORDS:  4

001   HW4-LIN1   (1,2,M     )  0          AMOUNT OF HOMEWORK (LINEAR):  DON'T HAVE, DON'T DO, MISSING
002   HW4-LIN2   (3         )  1          AMOUNT OF HOMEWORK (LINEAR):  ONE HALF HOUR
003   HW4-LIN3   (4         )  2          AMOUNT OF HOMEWORK (LINEAR):  ONE HOUR
004   HW4-LIN4   (5         )  3          AMOUNT OF HOMEWORK (LINEAR):  MORE THAN ONE HOUR

CONDITIONING ID:          BACK0014
DESCRIPTION:              AMOUNT OF HOMEWORK (QUADRATIC) (GRADE 4)
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              HMWRKQ4         LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                  B006601         DEGREES OF FREEDOM PER CONTRAST:  3
TYPE OF CONTRAST:         SCALE           NUMBER OF SPECIFICATION RECORDS:  4

001   HW4QUAD1   (1,2,M     )  0          AMOUNT OF HOMEWORK (QUAD):  DON'T HAVE, DON'T DO, MISSING
002   HW4QUAD2   (3         )  1          AMOUNT OF HOMEWORK (QUADRATIC):  ONE HALF HOUR
```

```
003   HW4QUAD3   (4           )  4          AMOUNT OF HOMEWORK (QUADRATIC):  ONE HOUR
004   HW4QUAD4   (5           )  9          AMOUNT OF HOMEWORK (QUADRATIC):  MORE THAN ONE HOUR

CONDITIONING ID:            BACK0018
DESCRIPTION:                PERCENT WHITE STUDENTS IN SCHOOL
GRADES/ASSESSMENTS:         N04, S04, N08, S08, N12
GROUP LABEL:                %WHITE              LENGTH OF CONTRAST FIELD       :   2
NAEP ID:                    PCTWHT              DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:   3

001   PREDOM/?   (80-110,M    )  00         PREDOMINANTLY WHITE, MISSING
002   MINORITY   (0-49        )  10         WHITE MINORITY
003   INTEGRAT   (50-79       )  01         INTEGRATED

CONDITIONING ID:            BACK0021
DESCRIPTION:                SINGLE/MULTIPLE PARENT(S) AT HOME
GRADES/ASSESSMENTS:         N04, S04, N08, S08, N12
GROUP LABEL:                PARENTS             LENGTH OF CONTRAST FIELD       :   1
NAEP ID:                    SINGLEP             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:   2

001   NOT2PARS   (2-4,M       )  0          NOT TWO PARENTS, MISSING
002   2PARENTS   (1           )  1          BOTH FATHER AND MOTHER AT HOME

CONDITIONING ID:            BACK0022
DESCRIPTION:                MOTHER AT HOME
GRADES/ASSESSMENTS:         N04, S04, N08, S08, N12
GROUP LABEL:                MOM@HOME            LENGTH OF CONTRAST FIELD       :   1
NAEP ID:                    B005601             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:   2

001   MOM@HM-N   (2,M         )  0          MOTHER AT HOME:  NO, MISSING
002   MOM@HM-Y   (1           )  1          MOTHER AT HOME:  YES

CONDITIONING ID:            BACK0023
DESCRIPTION:                PAGES READ FOR SCHOOL AND HOMEWORK EACH DAY (CONTRAST 1)
GRADES/ASSESSMENTS:         N04, S04, N08, S08, N12
GROUP LABEL:                PGSREAD1            LENGTH OF CONTRAST FIELD       :   1
NAEP ID:                    B001101             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:   2

001   <=5_PGS    (5,M         )  0          PAGES READ (1): 5 OR FEWER PAGES, MISSING
002   >=6_PGS    (1-4         )  1          PAGES READ (1): > 20 PGS, 16-20 PGS, 11-15 PGS, 6-10 PGS

CONDITIONING ID:            BACK0024
DESCRIPTION:                PAGES READ FOR SCHOOL AND HOMEWORK EACH DAY (CONTRAST 2)
GRADES/ASSESSMENTS:         N04, S04, N08, S08, N12
GROUP LABEL:                PGSREAD2            LENGTH OF CONTRAST FIELD       :   1
NAEP ID:                    B001101             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:   2

001   <=10_PGS   (4,5,M       )  0          PAGES READ (2):  6-10 PAGES, 5 OR FEWER PAGES, MISSING
002   >=11_PGS   (1-3         )  1          PAGES READ (2):  > 20 PAGES, 16-20 PAGES, 11-15 PAGES

CONDITIONING ID:            BACK0025
DESCRIPTION:                WENT TO PRESCHOOL?
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                PRESCH              LENGTH OF CONTRAST FIELD       :   1
NAEP ID:                    B004201             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:   2

001   PRESCH-N   (2,3,IDK,M   )  0          WENT TO PRESCHOOL?:  NO, I DON'T KNOW, MISSING
002   PRESCH-Y   (1           )  1          WENT TO PRESCHOOL?:  YES

CONDITIONING ID:            BACK0042
DESCRIPTION:                BORN IN ONE OF THE 50 STATES
GRADES/ASSESSMENTS:         N04, S04, N08, S08, N12
GROUP LABEL:                BORN_USA            LENGTH OF CONTRAST FIELD       :   1
NAEP ID:                    B007801             DEGREES OF FREEDOM PER CONTRAST:   1
```

```
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:   2

001   USA-YES    (1          )  0             BORN IN THE USA:  YES
002   USA-NO/?   (2,M        )  1             BORN IN THE USA:  NO/MIS SING

CONDITIONING ID:           BACK0043
DESCRIPTION:               HOW MANY TIMES CHANGED SCHOOLS IN THE LAST TWO YEARS?
GRADES/ASSESSMENTS:        N04, S04, N08, S08
GROUP LABEL:               SCH_CHGS           LENGTH OF CONTRAST FIELD        :   3
NAEP ID:                   B007301            DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:   4

001   CHGSCH=0   (1          )  000           CHANGED SCHOOLS (NONE)
002   CHGSCH=1   (2          )  100           CHANGED SCHOOLS ONCE
003   CHGSCH=2   (3          )  010           CHANGED SCHOOLS TWICE
004   CHGSCH3+   (4,M        )  001           CHANGED SCHOOLS 3 OR MORE TIMES, MISSING

CONDITIONING ID:           BACK0044
DESCRIPTION:               HOW MANY GRADES HAVE YOU GONE TO SCHOOL IN THIS STATE? (K-4)
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               GRDS_ST4           LENGTH OF CONTRAST FIELD        :   2
NAEP ID:                   B007601            DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:   3

001   ST4GRD<1   (1,M        )  00            LESS THAN ONE GRADE IN THIS STATE, MISSING (K-4)
002   ST4GRD12   (2          )  10            ONE TO TWO GRADES IN THIS STATE (K-4)
003   ST4GRD3+   (3          )  01            THREE OR MORE GRADES IN THIS STATE (K-4)

CONDITIONING ID:           BACK0045
DESCRIPTION:               HOW OFTEN DO YOU DISCUSS THINGS STUDIED IN SCHOOL WITH SOMEONE AT HOME?
GRADES/ASSESSMENTS:        N04, S04, N08, S08, N12
GROUP LABEL:               DISC@HOM           LENGTH OF CONTRAST FIELD        :   3
NAEP ID:                   B007401            DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:   4

001   DIS@HOM1   (1          )  000           DISCUSS AT HOME (ALMOST EVERYDAY)
002   DIS@HOM2   (2          )  100           DISCUSS AT HOME (ONCE OR TWICE A WEEK)
003   DIS@HOM3   (3          )  010           DISCUSS AT HOME (ONCE OR TWICE A MONTH)
004   DIS@HOM4   (4,M        )  001           DISCUSS AT HOME (NEVER OR HARDLY EVER, MISSING)

CONDITIONING ID:           BACK0046
DESCRIPTION:               HOW OFTEN DO USE A COMPUTER FOR SCHOOLWORK?
GRADES/ASSESSMENTS:        N04, S04, N08, S08, N12
GROUP LABEL:               COMP4SCH           LENGTH OF CONTRAST FIELD        :   4
NAEP ID:                   B007501            DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:   5

001   CMP4SCH1   (1          )  0000          COMPUTER FOR SCHOOLWORK (ALMOST EVERYDAY)
002   CMP4SCH2   (2          )  1000          COMPUTER FOR SCHOOLWORK (ONCE OR TWICE A WEEK)
003   CMP4SCH3   (3          )  0100          COMPUTER FOR SCHOOLWORK (ONCE OR TWICE A MONTH)
004   CMP4SCH4   (4          )  0010          COMPUTER FOR SCHOOLWORK (NEVER OR HARDLY EVER)
005   CMP4SCH5   (M          )  0001          COMPUTER FOR SCHOOLWORK (MISSING)

CONDITIONING ID:           READ0001
DESCRIPTION:               SCHOOL LEVEL AVERAGE READING PROFICIENCY
GRADES/ASSESSMENTS:        N04, S04, N08, S08, N12
GROUP LABEL:               SLP_READ           LENGTH OF CONTRAST FIELD        :   1
NAEP ID:                   SCHREAD            DEGREES OF FREEDOM PER CONTRAST: 999
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:   2

001   SLP_RD-Y   (@          )  1             SCHOOL LEVEL AVERAGE READING PROFICIENCY NOT-MISSING
002   SLP_RD-?   (M          )  0             SCHOOL LEVEL AVERAGE READING PROFICIENCY MISSING

CONDITIONING ID:           READ0002
DESCRIPTION:               SCHOOL LEVEL AVERAGE READING PROFICIENCY
GRADES/ASSESSMENTS:        N04, S04, N08, S08, N12
GROUP LABEL:               SLP_REDL           LENGTH OF CONTRAST FIELD        :   8
NAEP ID:                   SCHREAD            DEGREES OF FREEDOM PER CONTRAST: 999
TYPE OF CONTRAST:          SCALE              NUMBER OF SPECIFICATION RECORDS:   2
```

```
001   SLP_RD-L   (#           )  (F8.4)      SCHOOL LEVEL AVERAGE READING PROFICIENCY MEAN
002   SLP_RD-L   (M           )  0           SCHOOL LEVEL AVERAGE READING PROFICIENCY MISSING

CONDITIONING ID:            READ0003
DESCRIPTION:                DURING THE PAST MONTH, HOW MANY BOOKS HAVE YOU READ OUTSIDE OF SCHOOL?
GRADES/ASSESSMENTS:         N04, S04, N08, N12
GROUP LABEL:                NBOOKSRD          LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    R810801           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  5

001   NBOOKS-1   (1           )  0000        NUMBER OF BOOKS READ:  NONE
002   NBOOKS-2   (2           )  1000        NUMBER OF BOOKS READ:  ONE OR TWO
003   NBOOKS-3   (3           )  0100        NUMBER OF BOOKS READ:  THREE OR FOUR
004   NBOOKS-4   (4           )  0010        NUMBER OF BOOKS READ:  FIVE OR MORE
005   NBOOKS-?   (M           )  0001        NUMBER OF BOOKS READ:  MISSING

CONDITIONING ID:            READ0004
DESCRIPTION:                WHAT KIND OF READER DO YOU THINK YOU ARE?
GRADES/ASSESSMENTS:         N04, S04, N08, N12
GROUP LABEL:                KIND_RDR          LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    R810201           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  5

001   READ-VGD   (1           )  0000        KIND OF READER:  A VERY GOOD READER
002   READ-GD    (2           )  1000        KIND OF READER:  A GOOD READER
003   READ-AVG   (3           )  0100        KIND OF READER:  AN AVERAGE READER
004   READ-PR    (4           )  0010        KIND OF READER:  A POOR READER
005   READ-?     (M           )  0001        KIND OF READER:  MISSING

CONDITIONING ID:            READ0005
DESCRIPTION:                HOW OFTEN DO YOU READ FOR FUN ON YOUR OWN TIME?
GRADES/ASSESSMENTS:         N04, S04, N08, N12
GROUP LABEL:                READ4FUN          LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    R810901           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  5

001   RD4FUN-1   (1           )  0000        READ FOR FUN:  ALMOST EVERY DAY
002   RD4FUN-2   (2           )  1000        READ FOR FUN:  ONCE OR TWICE A WEEK
003   RD4FUN-3   (3           )  0100        READ FOR FUN:  ONCE OR TWICE A MONTH
004   RD4FUN-4   (4           )  0010        READ FOR FUN:  NEVER OR HARDLY EVER
005   RD4FUN-?   (M           )  0001        READ FOR FUN:  MISSING

CONDITIONING ID:            READ0009
DESCRIPTION:                HOW OFTEN DO YOU TALK WITH FRIENDS OR FAMILY ABOUT SOMTHING YOU HAVE READ?
GRADES/ASSESSMENTS:         N04, S04, N08, N12
GROUP LABEL:                TALKREAD          LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    R810902           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  5

001   TALKRD-1   (1           )  0000        TALK ABOUT READING: ALMOST EVERY DAY
002   TALKRD-2   (2           )  1000        TALK ABOUT READING: ONCE OR TWICE A WEEK
003   TALKRD-3   (3           )  0100        TALK ABOUT READING: ONCE OR TWICE A MONTH
004   TALKRD-4   (4           )  0010        TALK ABOUT READING: NEVER OR HARDLY EVER
005   TALKRD-?   (M           )  0001        TALK ABOUT READING: MISSING

CONDITIONING ID:            READ0010
DESCRIPTION:                HOW OFTEN DO YOU TAKE BOOKS OUT OF THE LIBRARY FOR YOU OWN ENJOYMENT?
GRADES/ASSESSMENTS:         N04, S04, N08, N12
GROUP LABEL:                USELIBRY          LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    R810903           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  5

001   USELIB-1   (1           )  0000        USE THE LIBRARY:  ALMOST EVERY DAY
002   USELIB-2   (2           )  1000        USE THE LIBRARY:  ONCE OR TWICE A WEEK
003   USELIB-3   (3           )  0100        USE THE LIBRARY:  ONCE OR TWICE A MONTH
004   USELIB-4   (4           )  0010        USE THE LIBRARY:  NEVER OR HARDLY EVER
005   USELIB-?   (M           )  0001        USE THE LIBRARY:  MISSING

CONDITIONING ID:            READ0011
```

```
DESCRIPTION:             HOW OFTEN DOES YOUR TEACHER DISCUSS NEW OR DIFFICULT VOCABULARY?
GRADES/ASSESSMENTS:      N04, S04, N08, N12
GROUP LABEL:             S_VOCAB            LENGTH OF CONTRAST FIELD      :  4
NAEP ID:                 R811001            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:        CLASS              NUMBER OF SPECIFICATION RECORDS:  5

001   VOCAB-S1   (1        )  0000         DISCUSS VOCABULARY:  ALMOST EVERY DAY
002   VOCAB-S2   (2        )  1000         DISCUSS VOCABULARY:  ONCE OR TWICE A WEEK
003   VOCAB-S3   (3        )  0100         DISCUSS VOCABULARY:  ONCE OR TWICE A MONTH
004   VOCAB-S4   (4        )  0010         DISCUSS VOCABULARY:  NEVER OR HARDLY EVER
005   VOCAB-S?   (M        )  0001         DISCUSS VOCABULARY:  MISSING

CONDITIONING ID:         READ0012
DESCRIPTION:             HOW OFTEN DOES TEACHER ASK STUDENTS TO TALK TO EACH OTHER ABOUT WHAT THEY READ?
GRADES/ASSESSMENTS:      N04, S04, N08, N12
GROUP LABEL:             S_TALKRD           LENGTH OF CONTRAST FIELD      :  4
NAEP ID:                 R811002            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:        CLASS              NUMBER OF SPECIFICATION RECORDS:  5

001   TLKRD-S1   (1        )  0000         TEACHER ASK TO TALK ABOUT READING:  ALMOST EVERY DAY
002   TLKRD-S2   (2        )  1000         TEACHER ASK TO TALK ABOUT READING:  ONCE OR TWICE A WEEK
003   TLKRD-S3   (3        )  0100         TEACHER ASK TO TALK ABOUT READING:  ONCE OR TWICE A MONTH
004   TLKRD-S4   (4        )  0010         TEACHER ASK TO TALK ABOUT READING:  NEVER OR HARDLY EVER
005   TLKRD-S?   (M        )  0001         TEACHER ASK TO TALK ABOUT READING:  MISSING

CONDITIONING ID:         READ0013
DESCRIPTION:             HOW OFTEN DOES TEACHER ASK YOU TO WORK IN A READING WORKBOOK OR ON A WORKSHEET?
GRADES/ASSESSMENTS:      N04, S04, N08, N12
GROUP LABEL:             S_WBKWSH           LENGTH OF CONTRAST FIELD      :  4
NAEP ID:                 R811003            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:        CLASS              NUMBER OF SPECIFICATION RECORDS:  5

001   WB/WS-S1   (1        )  0000         READING WORKBOOK/WORKSHEET:  ALMOST EVERY DAY
002   WB/WS-S2   (2        )  1000         READING WORKBOOK/WORKSHEET:  ONCE OR TWICE A WEEK
003   WB/WS-S3   (3        )  0100         READING WORKBOOK/WORKSHEET:  ONCE OR TWICE A MONTH
004   WB/WS-S4   (4        )  0010         READING WORKBOOK/WORKSHEET:  NEVER OR HARDLY EVER
005   WB/WS-S?   (M        )  0001         READING WORKBOOK/WORKSHEET:  MISSING

CONDITIONING ID:         READ0014
DESCRIPTION:             HOW OFTEN DOES YOUR TEACHER ASK YOU TO WRITE SOMETHING ABOUT WHAT YOU HAVE READ?
GRADES/ASSESSMENTS:      N04, S04, N08, N12
GROUP LABEL:             S_WRITRD           LENGTH OF CONTRAST FIELD      :  4
NAEP ID:                 R811004            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:        CLASS              NUMBER OF SPECIFICATION RECORDS:  5

001   WRTRD-S1   (1        )  0000         WRITE ABOUT READING:  ALMOST EVERY DAY
002   WRTRD-S2   (2        )  1000         WRITE ABOUT READING:  ONCE OR TWICE A WEEK
003   WRTRD-S3   (3        )  0100         WRITE ABOUT READING:  ONCE OR TWICE A MONTH
004   WRTRD-S4   (4        )  0010         WRITE ABOUT READING:  NEVER OR HARDLY EVER
005   WRTRD-S?   (M        )  0001         WRITE ABOUT READING:  MISSING

CONDITIONING ID:         READ0015
DESCRIPTION:             HOW OFTEN DOES YOUR TEACHER ASK TO DO GROUP ACTIVITY/PROJECT ABOUT WHAT IS READ?
GRADES/ASSESSMENTS:      N04, S04, N08, N12
GROUP LABEL:             S_RDPROJ           LENGTH OF CONTRAST FIELD      :  4
NAEP ID:                 R811005            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:        CLASS              NUMBER OF SPECIFICATION RECORDS:  5

001   RDPRJ-S1   (1        )  0000         PROJECT ABOUT READING: ALMOST EVERY DAY
002   RDPRJ-S2   (2        )  1000         PROJECT ABOUT READING: ONCE OR TWICE A WEEK
003   RDPRJ-S3   (3        )  0100         PROJECT ABOUT READING: ONCE OR TWICE A MONTH
004   RDPRJ-S4   (4        )  0010         PROJECT ABOUT READING: NEVER OR HARDLY EVER
005   RDPRJ-S?   (M        )  0001         PROJECT ABOUT READING: MISSING

CONDITIONING ID:         READ0016
DESCRIPTION:             HOW OFTEN DOES YOUR TEACHER ASK STUDENTS TO READ ALOUD?
GRADES/ASSESSMENTS:      N04, S04, N08, N12
GROUP LABEL:             S_ALOUD            LENGTH OF CONTRAST FIELD      :  4
NAEP ID:                 R811006            DEGREES OF FREEDOM PER CONTRAST:  1
```

```
TYPE OF CONTRAST:         CLASS              NUMBER OF SPECIFICATION RECORDS:   5

001   ALOUD-S1   (1         )  0000         READ ALOUD: ALMOST EVERY DAY
002   ALOUD-S2   (2         )  1000         READ ALOUD: ONCE OR TWICE A WEEK
003   ALOUD-S3   (3         )  0100         READ ALOUD: ONCE OR TWICE A MONTH
004   ALOUD-S4   (4         )  0010         READ ALOUD: NEVER OR HARDLY EVER
005   ALOUD-S?   (M         )  0001         READ ALOUD: MISSING

CONDITIONING ID:          READ0017
DESCRIPTION:              HOW OFTEN DOES YOUR TEACHER ASK YOU TO READ SILENTLY?
GRADES/ASSESSMENTS:       N04, S04, N08, N12
GROUP LABEL:              S_SILENT           LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                  R811007            DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS              NUMBER OF SPECIFICATION RECORDS:   5

001   SILNT-S1   (1         )  0000         READ SILENTLY: ALMOST EVERY DAY
002   SILNT-S2   (2         )  1000         READ SILENTLY: ONCE OR TWICE A WEEK
003   SILNT-S3   (3         )  0100         READ SILENTLY: ONCE OR TWICE A MONTH
004   SILNT-S4   (4         )  0010         READ SILENTLY: NEVER OR HARDLY EVER
005   SILNT-S?   (M         )  0001         READ SILENTLY: MISSING

CONDITIONING ID:          READ0018
DESCRIPTION:              HOW OFTEN DOES TEACHER ASK TO WRITE IN LOG OR JOURNAL ABOUT WHAT YOU HAVE READ?
GRADES/ASSESSMENTS:       N04, S04, N08, N12
GROUP LABEL:              S_RDLOG            LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                  R811008            DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS              NUMBER OF SPECIFICATION RECORDS:   5

001   RDLOG-S1   (1         )  0000         WRITE IN LOG/JOURNAL:  ALMOST EVERY DAY
002   RDLOG-S2   (2         )  1000         WRITE IN LOG/JOURNAL:  ONCE OR TWICE A WEEK
003   RDLOG-S3   (3         )  0100         WRITE IN LOG/JOURNAL:  ONCE OR TWICE A MONTH
004   RDLOG-S4   (4         )  0010         WRITE IN LOG/JOURNAL:  NEVER OR HARDLY EVER
005   RDLOG-S?   (M         )  0001         WRITE IN LOG/JOURNAL:  MISSING

CONDITIONING ID:          READ0019
DESCRIPTION:              HOW OFTEN DOES TEACHER GIVE YOU TIME TO READ BOOKS YOU HAVE CHOSEN YOURSELF?
GRADES/ASSESSMENTS:       N04, S04, N08, N12
GROUP LABEL:              S_OWNBKS           LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                  R811009            DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS              NUMBER OF SPECIFICATION RECORDS:   5

001   OWNBK-S1   (1         )  0000         BOOKS CHOSEN YOURSELF:  ALMOST EVERY DAY
002   OWNBK-S2   (2         )  1000         BOOKS CHOSEN YOURSELF:  OR TWICE A WEEK
003   OWNBK-S3   (3         )  0100         BOOKS CHOSEN YOURSELF:  ONCE OR TWICE A MONTH
004   OWNBK-S4   (4         )  0010         BOOKS CHOSEN YOURSELF:  NEVER OR HARDLY EVER
005   OWNBK-S?   (M         )  0001         BOOKS CHOSEN YOURSELF:  MISSING

CONDITIONING ID:          READ0027
DESCRIPTION:              ABOUT HOW MANY QUESTIONS DID YOU GET RIGHT ON THE READING TEST?
GRADES/ASSESSMENTS:       N04, S04, N08, N12
GROUP LABEL:              #QUESTN+           LENGTH OF CONTRAST FIELD       :   3
NAEP ID:                  RM00101            DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS              NUMBER OF SPECIFICATION RECORDS:   4

001   #QUEST+1   (1         )  000          NUMBER QUESTIONS RIGHT:  ALMOST ALL
002   #QUEST+2   (2         )  100          NUMBER QUESTIONS RIGHT:  MORE THAN HALF
003   #QUEST+3   (3         )  010          NUMBER QUESTIONS RIGHT:  ABOUT HALF
004   #QUEST+4   (4,M       )  001          NUMBER QUESTIONS RIGHT:  LESS THAN HALF, MISSING

CONDITIONING ID:          READ0028
DESCRIPTION:              HOW HARD WAS THIS READING TEST COMPARED TO OTHERS?
GRADES/ASSESSMENTS:       N04, S04, N08, N12
GROUP LABEL:              TEST_DIF           LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                  RM00201            DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:         CLASS              NUMBER OF SPECIFICATION RECORDS:   5

001   TESTDIF1   (1         )  0000         TEST DIFFICULTY:  MUCH HARDER THAN OTHERS
002   TESTDIF2   (2         )  1000         TEST DIFFICULTY:  HARDER THAN OTHERS
003   TESTDIF3   (3         )  0100         TEST DIFFICULTY:  ABOUT AS HARD AS OTHERS
```

```
004   TESTDIF4   (4          )  0010        TEST DIFFICULTY:  EASIER THAN OTHERS
005   TESTDIF?   (M          )  0001        TEST DIFFICULTY:  MISSING

CONDITIONING ID:             READ0029
DESCRIPTION:                 HOW HARD DID YOU TRY ON THIS TEST COMPARED TO OTHER READING TESTS?
GRADES/ASSESSMENTS:          N04, S04, N08, N12
GROUP LABEL:                 TEST_EFF               LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                     RM00301                DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:            CLASS                  NUMBER OF SPECIFICATION RECORDS:   5

001   TESTEFF1   (1          )  0000        TEST EFFORT:  MUCH HARDER THAN OTHERS
002   TESTEFF2   (2          )  1000        TEST EFFORT:  HARDER THAN OTHERS
003   TESTEFF3   (3          )  0100        TEST EFFORT:  ABOUT AS HARD AS OTHERS
004   TESTEFF4   (4          )  0010        TEST EFFORT:  NOT AS HARD AS OTHERS
005   TESTEFF?   (M          )  0001        TEST EFFORT:  MISSING

CONDITIONING ID:             READ0030
DESCRIPTION:                 HOW IMPORTANT WAS IT TO YOU TO DO WELL ON THE READING TEST?
GRADES/ASSESSMENTS:          N04, S04, N08, N12
GROUP LABEL:                 TEST_IMP               LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                     RM00401                DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:            CLASS                  NUMBER OF SPECIFICATION RECORDS:   5

001   TESTIMP1   (1          )  0000        TEST IMPORTANCE:  VERY IMPORTANT
002   TESTIMP2   (2          )  1000        TEST IMPORTANCE:  IMPORTANT
003   TESTIMP3   (3          )  0100        TEST IMPORTANCE:  SOMEWHAT IMPORTANT
004   TESTIMP4   (4          )  0010        TEST IMPORTANCE:  NOT VERY IMPORTANT
005   TESTIMP?   (M          )  0001        TEST IMPORTANCE:  MISSING

CONDITIONING ID:             READ0031
DESCRIPTION:                 HOW OFTEN WERE YOU ASKED TO WRITE LONG ANSWERS ON READING TESTS?
GRADES/ASSESSMENTS:          N04, S04, N08, N12
GROUP LABEL:                 LONG_ANS               LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                     RM00501                DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:            CLASS                  NUMBER OF SPECIFICATION RECORDS:   5

001   DSOLUTN1   (1          )  0000        DETAILED SOLUTIONS:  AT LEAST ONCE A WEEK
002   DSOLUTN2   (2          )  1000        DETAILED SOLUTIONS:  ONCE OR TWICE A MONTH
003   DSOLUTN3   (3          )  0100        DETAILED SOLUTIONS:  ONCE OR TWICE A YEAR
004   DSOLUTN4   (4          )  0010        DETAILED SOLUTIONS:  NEVER
005   DSOLUTN5   (M          )  0001        DETAILED SOLUTIONS:  MISSING

CONDITIONING ID:             SCHL0002
DESCRIPTION:                 HAS READING BEEN IDENTIFIED AS A PRIORITY? (GRADE 4)
GRADES/ASSESSMENTS:          N04, S04, N08, S08
GROUP LABEL:                 PRIOR-RD               LENGTH OF CONTRAST FIELD       :   2
NAEP ID:                     C031601                DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:            CLASS                  NUMBER OF SPECIFICATION RECORDS:   3

001   RPRIOR-Y   (1          )  00          READING PRIORITY:  YES
002   RPRIOR-N   (2          )  10          READING PRIORITY:  NO
003   RPRIOR-?   (M          )  01          READING PRIORITY:  MISSING

CONDITIONING ID:             SCHL0003
DESCRIPTION:                 HAS WRITING BEEN IDENTIFIED AS A PRIORITY? (GRADE 4)
GRADES/ASSESSMENTS:          N04, S04, N08, S08
GROUP LABEL:                 PRIOR-WR               LENGTH OF CONTRAST FIELD       :   2
NAEP ID:                     C031602                DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:            CLASS                  NUMBER OF SPECIFICATION RECORDS:   3

001   WPRIOR-Y   (1          )  00          WRITING PRIORITY:  YES
002   WPRIOR-N   (2          )  10          WRITING PRIORITY:  NO
003   WPRIOR-?   (M          )  01          WRITING PRIORITY:  MISSING

CONDITIONING ID:             SCHL0004
DESCRIPTION:                 WHAT PERCENT OF STUDENTS RECEIVE SUBSIDIZED LUNCH?
GRADES/ASSESSMENTS:          N04, S04, N08, S08, N12
GROUP LABEL:                 %SUBLUN                LENGTH OF CONTRAST FIELD       :   5
NAEP ID:                     C032001                DEGREES OF FREEDOM PER CONTRAST:   1
```

```
TYPE OF CONTRAST:          CLASS               NUMBER OF SPECIFICATION RECORDS:   6

001   %SUBLUN1   (1,2,3       )  00000         PERCENT SUBSIDIZED LUNCH:  NONE-10%
002   %SUBLUN2   (4           )  10000         PERCENT SUBSIDIZED LUNCH:  11-25%
003   %SUBLUN3   (5           )  01000         PERCENT SUBSIDIZED LUNCH:  26-50%
004   %SUBLUN4   (6           )  00100         PERCENT SUBSIDIZED LUNCH:  51-75%
005   %SUBLUN5   (7,8         )  00010         PERCENT SUBSIDIZED LUNCH:  76-100%
006   %SUBLUN?   (M           )  00001         PERCENT SUBSIDIZED LUNCH:  MISSING

CONDITIONING ID:           SCHL0005
DESCRIPTION:               WHAT PERCENT OF STUDENTS RECEIVE REMEDIAL READING?
GRADES/ASSESSMENTS:        N04, S04, N08, S08, N12
GROUP LABEL:               %REMDL-R            LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                   C032002             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:          CLASS               NUMBER OF SPECIFICATION RECORDS:   5

001   %REMRED1   (1,2         )  0000          PERCENT REMEDIAL READING:  NONE-5%
002   %REMRED2   (3           )  1000          PERCENT REMEDIAL READING:  6-10%
003   %REMRED3   (4           )  0100          PERCENT REMEDIAL READING:  11-25%
004   %REMRED4   (5,6,7,8     )  0010          PERCENT REMEDIAL READING:  26-100%
005   %REMRED?   (M           )  0001          PERCENT REMEDIAL READING:  MISSING

CONDITIONING ID:           SCHL0006
DESCRIPTION:               WHAT PERCENTAGE OF STUDENTS ARE ENROLLED AT BEGINNING AND END OF SCHOOL YEAR?
GRADES/ASSESSMENTS:        N04, S04, N08, S08, N12
GROUP LABEL:               %ENR/YR             LENGTH OF CONTRAST FIELD       :   4
NAEP ID:                   C033700             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:          CLASS               NUMBER OF SPECIFICATION RECORDS:   5

001   %ENR/YR1   (1           )  0000          YEAR LONG ENROLLMENT:  98-100 PERCENT
002   %ENR/YR2   (2           )  1000          YEAR LONG ENROLLMENT:  95-97 PERCENT
003   %ENR/YR3   (3           )  0100          YEAR LONG ENROLLMENT:  90-94 PERCENT
004   %ENR/YR4   (4           )  0010          YEAR LONG ENROLLMENT:  LESS THAN 90 PERCENT
005   %ENR/YR?   (M           )  0001          YEAR LONG ENROLLMENT:  MISSING

CONDITIONING ID:           SCHL0007
DESCRIPTION:               HOW IS 4TH GRADE ORGANIZED AT YOUR SCHOOL?
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               ORGANIZ4            LENGTH OF CONTRAST FIELD       :   3
NAEP ID:                   C030900             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:          CLASS               NUMBER OF SPECIFICATION RECORDS:   4

001   SELFCONT   (1           )  000           4TH GRADE ORGANIZATION:  SELF CONTAINED
002   DEPTLIZD   (2           )  100           4TH GRADE ORGANIZATION:  DEPARTMENTALIZED
003   REGRPED    (3           )  010           4TH GRADE ORGANIZATION:  REGROUPED
004   ORGANIZ?   (M           )  001           4TH GRADE ORGANIZATION:  MISSING

CONDITIONING ID:           SCHL0009
DESCRIPTION:               ARE 4TH GRADERS ASSIGNED TO CLASSES BY ABILITY?
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               CLASS/AB            LENGTH OF CONTRAST FIELD       :   2
NAEP ID:                   C031100             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:          CLASS               NUMBER OF SPECIFICATION RECORDS:   3

001   ABILTY-Y   (1           )  00            4TH GRADERS ASSIGNED BY ABILITY:  YES
002   ABILTY-N   (2           )  10            4TH GRADERS ASSIGNED BY ABILITY:  NO
003   ABILTY-?   (M           )  01            4TH GRADERS ASSIGNED BY ABILITY:  MISSING

CONDITIONING ID:           SCHL0014
DESCRIPTION:               POLICY CONTROLLING TIME FOR READING INSTRUCTION?
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               POLICY-R            LENGTH OF CONTRAST FIELD       :   2
NAEP ID:                   C031301             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:          CLASS               NUMBER OF SPECIFICATION RECORDS:   3

001   RD_POL-Y   (1           )  00            READING TIME POLICY:  YES
002   RD_POL-N   (2           )  10            READING TIME POLICY:  NO
003   RD_POL-?   (M           )  01            READING TIME POLICY:  MISSING
```

```
CONDITIONING ID:            SCHL0015
DESCRIPTION:                POLICY CONTROLLING TIME FOR WRITING INSTRUCTION?
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                POLICY-W          LENGTH OF CONTRAST FIELD      :  2
NAEP ID:                    C031302           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  3

001   WR_POL-Y   (1          )  00          WRITING TIME POLICY:  YES
002   WR_POL-N   (2          )  10          WRITING TIME POLICY:  NO
003   WR_POL-?   (M          )  01          WRITING TIME POLICY:  MISSING

CONDITIONING ID:            SCHL0020
DESCRIPTION:                DOES SCHOOL INVOLVE PARENTS AS AIDES IN CLASS?
GRADES/ASSESSMENTS:         N04, S04, N08, S08, N12
GROUP LABEL:                PAR_AIDE          LENGTH OF CONTRAST FIELD      :  3
NAEP ID:                    C032207           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  4

001   PARAID-R   (1          )  000         PARENTS AS AIDES IN CLASS:  ROUTINELY
002   PARAID-O   (2          )  100         PARENTS AS AIDES IN CLASS:  OCCASIONALLY
003   PARAID-N   (3          )  010         PARENTS AS AIDES IN CLASS:  NO
004   PARAID-?   (M          )  001         PARENTS AS AIDES IN CLASS:  MISSING

CONDITIONING ID:            TCHR0001
DESCRIPTION:                HOW WELL DOES SCHOOL PROVIDE RESOURCES
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                RESOURCE          LENGTH OF CONTRAST FIELD      :  4
NAEP ID:                    T041201           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  5

001   RESOURC1   (1          )  0000        RESOURCES:  GET ALL
002   RESOURC2   (2          )  1000        RESOURCES:  GET MOST
003   RESOURC3   (3          )  0100        RESOURCES:  GET SOME
004   RESOURC4   (4          )  0010        RESOURCES:  DON'T GET
005   RESOURC?   (M,DNA      )  0001        RESOURCES:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TCHR0002
DESCRIPTION:                TEACHER MATCH STATUS WITH STUDENT
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                T_MATCH           LENGTH OF CONTRAST FIELD      :  2
NAEP ID:                    TCHMTCH           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  3

001   TMCH-NO    (1,M        )  00          TEACHER MATCH:  NO MATCH
002   TMCH-PAR   (2          )  10          TEACHER MATCH:  PARTIAL MATCH
003   TMCH-COM   (3          )  01          TEACHER MATCH:  COMPLETE MATCH

CONDITIONING ID:            TCHR0003
DESCRIPTION:                TEACHER GENDER
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                T_GENDER          LENGTH OF CONTRAST FIELD      :  2
NAEP ID:                    T040001           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  3

001   T_MALE     (1          )  00          TEACHER GENDER:  MALE
002   T_FEMALE   (2          )  10          TEACHER GENDER:  FEMALE
003   T_SEX-?    (M,DNA      )  01          TEACHER GENDER:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TCHR0004
DESCRIPTION:                TEACHER RACE/ETHNICITY
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                T_RACE            LENGTH OF CONTRAST FIELD      :  5
NAEP ID:                    T040101           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS             NUMBER OF SPECIFICATION RECORDS:  6

001   T_WHITE    (1          )  00000       TEACHER ETHNICITY:  WHITE
002   T_BLACK    (2          )  10000       TEACHER ETHNICITY:  BLACK
003   T_HISP     (3          )  01000       TEACHER ETHNICITY:  HISPANIC
004   T_ASIAN    (4          )  00100       TEACHER ETHNICITY:  ASIAN, PACIFIC ISLANDER
```

```
005   T_AM.IND   (5          )  00010       TEACHER ETHNICITY:  AMERICAN INDIAN, ALASKAN NATIVE
006   T_RACE-?   (M,DNA      )  00001       TEACHER ETHNICITY:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TCHR0005
DESCRIPTION:                TEACHER HISPANIC BACKGROUND
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                T_HISPBK            LENGTH OF CONTRAST FIELD       :  5
NAEP ID:                    T040201             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  6

001   T_NONHSP   (1          )  00000       TEACHER HISPANIC BACKGROUND:  NOT HISPANIC
002   T_MEXICN   (2          )  10000       TEACHER HISPANIC BACKGROUND:  MEXICAN/MEXICAN AMERICAN
003   T_PUERTO   (3          )  01000       TEACHER HISPANIC BACKGROUND:  PUERTO RICAN
004   T_CUBAN    (4          )  00100       TEACHER HISPANIC BACKGROUND:  CUBAN
005   T_OTHER    (5          )  00010       TEACHER HISPANIC BACKGROUND:  OTHER
006   T_HISP-?   (M,DNA      )  00001       TEACHER HISPANIC BACKGROUND:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TCHR0006
DESCRIPTION:                YEARS TEACHING ELEMENTARY/SECONDARY SCHOOL
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                T_YRSEXP            LENGTH OF CONTRAST FIELD       :  5
NAEP ID:                    T040301             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  6

001   T_YREXP1   (1          )  00000       YEARS TEACHING:  2 OR LESS YEARS
002   T_YREXP2   (2          )  10000       YEARS TEACHING:  3-5 YEARS
003   T_YREXP3   (3          )  01000       YEARS TEACHING:  6-10 YEARS
004   T_YREXP4   (4          )  00100       YEARS TEACHING:  11-24 YEARS
005   T_YREXP5   (5          )  00010       YEARS TEACHING:  25 OR MORE YEARS
006   T_YREXP?   (M,DNA      )  00001       YEARS TEACHING:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TCHR0007
DESCRIPTION:                TYPE OF TEACHING CERTIFICATION
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                TCH_CERT            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T040401             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  4

001   TCERT-NO   (1          )  000         TEACHING CERTIFICATION:  NONE, TEMPORARY, PROVISIONAL
002   TCERT-RG   (2          )  100         TEACHING CERTIFICATION:  REGULAR, NOT HIGHEST AVAILABLE
003   TCERT-HI   (3          )  010         TEACHING CERTIFICATION:  HIGHEST AVAILABLE
004   TCERT-?    (M,DNA      )  001         TEACHING CERTIFICATION:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TCHR0008
DESCRIPTION:                TEACHER GENERAL CERTIFICATION (ELEMENTARY, MIDDLE/JUNIOR HS EDUCATION)
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                CERT-GEN            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T040501             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  4

001   CERTG-Y    (1          )  000         GENERAL CERTIFICATION:  YES
002   CERTG-N    (2          )  100         GENERAL CERTIFICATION:  NO
003   CERTG-NS   (3          )  010         GENERAL CERTIFICATION:  NOT OFFERED IN STATE
004   CERTG-?    (M,DNA      )  001         GENERAL CERTIFICATION:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TCHR0009
DESCRIPTION:                TEACHER'S HIGHEST ACADEMIC DEGREE
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                T_DEGREE            LENGTH OF CONTRAST FIELD       :  6
NAEP ID:                    T040601             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  7

001   <BACHLRS   (1          )  000000      TEACHER DEGREE:  LESS THAN A BACHELOR'S DEGREE
002   BACHELRS   (2          )  100000      TEACHER DEGREE:  BACHELOR'S DEGREE
003   MASTERS    (3          )  010000      TEACHER DEGREE:  MASTER'S DEGREE
004   SPECLIST   ([illegible]          )  001000      TEACHER DEGREE:  EDUCATION SPECIALIST
005   DOCTORAT   (5          )  000100      TEACHER DEGREE:  DOCTORATE
006   PROFESSL   (6          )  000010      TEACHER DEGREE:  PROFESSIONAL DEGREE
007   DEGREE-?   (M,DNA      )  000001      TEACHER DEGREE:  MISSING, DOES NOT APPLY
```

```
CONDITIONING ID:            TCHR0010
DESCRIPTION:                TEACHER UNDERGRADUATE MAJOR IN EDUCATION
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                UGRAD_ED            LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                    T040701             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  2

001   UGR_ED-?   (M,DNA      )  0               UNDERGRADUATE EDUCATION MAJOR:  MISSING, DOES NOT APPLY
002   UGR_ED-Y   (1          )  1               UNDERGRADUATE EDUCATION MAJOR:  YES

CONDITIONING ID:            TCHR0011
DESCRIPTION:                TEACHER UNDERGRADUATE MAJOR IN ENGLISH/READING/LANGUAGE ARTS
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                UGRAD_RD            LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                    T040702             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  2

001   UGR_RD-?   (M,DNA      )  0               UNDERGRADUATE READING MAJOR:  MISSING, DOES NOT APPLY
002   UGR_RD-Y   (1          )  1               UNDERGRADUATE READING MAJOR:  YES

CONDITIONING ID:            TCHR0012
DESCRIPTION:                TEACHER GRADUATE MAJOR IN EDUCATION
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                GRAD_ED             LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                    T040801             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  2

001   GR_ED-?    (M,DNA      )  0               GRADUATE EDUCATION MAJOR:  MISSING, DOES NOT APPLY
002   GR_ED-Y    (1          )  1               GRADUATE EDUCATION MAJOR:  YES

CONDITIONING ID:            TCHR0013
DESCRIPTION:                TEACHER GRADUATE MAJOR IN ENGLISH/READING/LANGUAGE ARTS
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                GRAD_RD             LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                    T040802             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  2

001   GR_RD-?    (M,DNA      )  0               GRADUATE READING MAJOR:  MISSING, DOES NOT APPLY
002   GR_RD-Y    (1          )  1               GRADUATE READING MAJOR:  YES

CONDITIONING ID:            TCHR0014
DESCRIPTION:                NO TEACHER GRADUATE-LEVEL STUDY
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                GRAD_NO             LENGTH OF CONTRAST FIELD       :  1
NAEP ID:                    T040806             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  2

001   GR_NO-?    (M,DNA      )  0               NO GRAUDATE STUDY:  MISSING, DOES NOT APPLY
002   GR_NO-Y    (1          )  1               NO GRADUATE STUDY:  YES

CONDITIONING ID:            TCHR0015
DESCRIPTION:                ARE CURRICULUM SPECIALISTS AVAILABLE FOR READING?
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                CURSPE-R            LENGTH OF CONTRAST FIELD       :  2
NAEP ID:                    T041301             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  3

001   READCS-Y   (1          )  00              READING CURRICULUM SPECIALISTS:  YES
002   READCS-N   (2          )  10              READING CURRICULUM SPECIALISTS:  NO
003   READCS-?   (M,DNA      )  01              READING CURRICULUM SPECIALISTS:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TCHR0016
DESCRIPTION:                HOW OFTEN DO AIDES ASSIST YOU IN CLASS?
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                TCH_AIDE            LENGTH OF CONTRAST FIELD       :  5
NAEP ID:                    T041401             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  6
```

```
001   TCHAIDE1   (1          )  00000       TEACHER AIDES IN CLASS:  EVERY DAY
002   TCHAIDE2   (2          )  10000       TEACHER AIDES IN CLASS:  SEVERAL TIMES A WEEK
003   TCHAIDE3   (3          )  01000       TEACHER AIDES IN CLASS:  ONCE A WEEK
004   TCHAIDE4   (4          )  00100       TEACHER AIDES IN CLASS:  LESS THAN ONCE A WEEK
005   TCHAIDE5   (5          )  00010       TEACHER AIDES IN CLASS:  NEVER
006   TCHAIDE?   (M,DNA      )  00001       TEACHER AIDES IN CLASS:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TCHR0017
DESCRIPTION:                NUMBER OF STUDENTS IN CLASS
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                T_NCLASS            LENGTH OF CONTRAST FIELD       :   5
NAEP ID:                    TCHNCLS             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:   6

001   T_NCLAS1   (0-20       )  00000       CLASS SIZE:  0-20
002   T_NCLAS2   (21-25      )  10000       CLASS SIZE:  21-25
003   T_NCLAS3   (26-31      )  01000       CLASS SIZE:  26-30
004   T_NCLAS4   (31-35      )  00100       CLASS SIZE:  31-35
005   T_NCLAS5   (36-61      )  00010       CLASS SIZE:  36-60
006   T_NCLAS?   (M          )  00001       CLASS SIZE:  MISSING

CONDITIONING ID:            TRED0001
DESCRIPTION:                TEACHER HOURS SPENT IN IN-SERVICE READING EDUCATION
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                INSERV_R            LENGTH OF CONTRAST FIELD       :   5
NAEP ID:                    T041001             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:   6

001   INSERVR1   (1          )  00000       READING HOURS IN-SERVICE:  NONE
002   INSERVR2   (2          )  10000       READING HOURS IN-SERVICE:  LESS THAN 6 HOURS
003   INSERVR3   (3          )  01000       READING HOURS IN-SERVICE:  6-15 HOURS
004   INSERVR4   (4          )  00100       READING HOURS IN-SERVICE:  16-35 HOURS
005   INSERVR5   (5          )  00010       READING HOURS IN-SERVICE:  MORE THAN 35 HOURS
006   INSERVR?   (M,DNA      )  00001       READING HOURS IN-SERVICE:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0002
DESCRIPTION:                TEACHER CERTIFICATION IN READING
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                CERT-RED            LENGTH OF CONTRAST FIELD       :   3
NAEP ID:                    T040502             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:   4

001   CERTR-Y    (1          )  000         READING CERTIFICATION:  YES
002   CERTR-N    (2          )  100         READING CERTIFICATION:  NO
003   CERTR-NS   (3          )  010         READING CERTIFICATION:  NOT OFFERED IN STATE
004   CERTR-?    (M,DNA      )  001         READING CERTIFICATION:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0003
DESCRIPTION:                TEACHER CERTIFICATION MIDDLE/JUNIOR HS/SECONDARY ENGLISH/LANGUAGE ARTS
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                CERT-ENG            LENGTH OF CONTRAST FIELD       :   3
NAEP ID:                    T040503             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:   4

001   CERTE-Y    (1          )  000         READING CERTIFICATION:  YES
002   CERTE-N    (2          )  100         READING CERTIFICATION:  NO
003   CERTE-NS   (3          )  010         READING CERTIFICATION:  NOT OFFERED IN STATE
004   CERTE-?    (M,DNA      )  001         READING CERTIFICATION:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0004
DESCRIPTION:                ARE STUDENTS ASSIGNED TO READING CLASS BY ABILITY?
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                ABIL_CLA            LENGTH OF CONTRAST FIELD       :   2
NAEP ID:                    T046101             DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:   3

001   AB_CLA-Y   (1          )  00          READING CLASS BY ABILITY: YES
002   AB_CLA-N   (2          )  10          READING CLASS BY ABILITY: NO
003   AB_CLA-?   (M,DNA      )  01          READING CLASS BY ABILITY: MISSING, DOES NOT APPLY
```

```
CONDITIONING ID:            TRED0005
DESCRIPTION:                READING ABILITY LEVEL OF STUDENTS IN CLASS
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                ABIL_RED           LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    T046201            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS              NUMBER OF SPECIFICATION RECORDS:  5

001  AB_READ1  (1         )  0000        READING ABILITY:  PRIMARILY HIGH ABILITY
002  AB_READ2  (2         )  1000        READING ABILITY:  PRIMARILY AVERAGE ABILITY
003  AB_READ3  (3         )  0100        READING ABILITY:  PRIMARILY LOW ABILITY
004  AB_READ4  (4         )  0010        READING ABILITY:  WIDELY MIXED ABILITY
005  AB_READ?  (M,DNA     )  0001        READING ABILITY:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0006
DESCRIPTION:                TIME SPENT PER DAY ON READING INSTRUCTION
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                INS_TIME           LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T046301            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS              NUMBER OF SPECIFICATION RECORDS:  4

001   INSTIME1  (1,2      )  000         READING INSTRUCTION TIME:  30-45 MINUTES/DAY
002   INSTIME2  (3        )  100         READING INSTRUCTION TIME:  60 MINUTES/DAY
003   INSTIME3  (4        )  010         READING INSTRUCTION TIME:  90 MINUTES OR MORE/DAY
004   INSTIME?  (M,DNA    )  001         READING INSTRUCTION TIME:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0007
DESCRIPTION:                NUMBER OF INSTRUCTIONAL GROUPS CLASS DIVIDED INTO FOR READING
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                #INS_GRP           LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T046400            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS              NUMBER OF SPECIFICATION RECORDS:  4

001   INSGRP-1  (1        )  000         # INSTRUCTIONAL GROUPS:  WHOLE CLASS ACTIVITY
002   INSGRP-2  (2        )  100         # INSTRUCTIONAL GROUPS:  WHOLE CLASS/FLEXIBLE GROUPING
003   INSGRP-3  (3,4,5,6  )  010         # INSTRUCTIONAL GROUPS:  TWO OR MORE
004   INSGRP-4  (7,M,DNA  )  001         # INSTRUCTIONAL GROUPS:  INDIVID INSTRUC, MSSNG, DOESN'T APP

CONDITIONING ID:            TRED0008
DESCRIPTION:                TYPE OF MATERIALS FORMING CORE OF READING PROGRAM
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                RMATERLS           LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    T046501            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS              NUMBER OF SPECIFICATION RECORDS:  5

001   BASAL_RM  (1        )  0000        TYPE OF READING MATERIALS:  BASAL
002   TRADE_RM  (2        )  1000        TYPE OF READING MATERIALS:  TRADE
003   BAS&TRA   (3        )  0100        TYPE OF READING MATERIALS:  BASAL AND TRADE
004   OTHER_RM  (4        )  0010        TYPE OF READING MATERIALS:  OTHER
005   RMATS-?   (M,DNA    )  0001        TYPE OF READING MATERIALS:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0009
DESCRIPTION:                HOW OFTEN ARE NEWSPAPERS/MAGAZINES USED TO TEACH READING CLASS
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                NEWS/MAG           LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T046601            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS              NUMBER OF SPECIFICATION RECORDS:  4

001   NEWMAG-1  (1,2      )  000         NEWSPPRS/MAGAZNS (TEACHER): ALMOST EVERY DAY,ONCE/TWICE WEEK
002   NEWMAG-2  (3        )  100         NEWSPAPERS/MAGAZINES (TEACHER):  ONCE OR TWICE A MONTH
003   NEWMAG-3  (4        )  010         NEWSPAPERS/MAGAZINES (TEACHER):  NEVER OF HARDLEY EVER
004   NEWMAG-?  (M,DNA    )  001         NEWSPAPERS/MAGAZINES (TEACHER):  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0010
DESCRIPTION:                HOW OFTEN ARE READING KITS USED TO TEACH READING CLASS
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                READKITS           LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T046602            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS              NUMBER OF SPECIFICATION RECORDS:  4
```

```
001   RDKITS-1   (1,2         )  000         READING KITS (TEACHER):  ALMOST EVERY DAY, ONCE/TWICE A WEEK
002   RDKITS-2   (3           )  100         READING KITS (TEACHER):  ONCE OR TWICE A MONTH
003   RDKITS-3   (4           )  010         READING KITS (TEACHER):  NEVER OF HARDLEY EVER
004   RDKITS-?   (M,DNA       )  001         READING KITS (TEACHER):  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0011
DESCRIPTION:                HOW OFTEN IS READING COMPUTER SOFTWARE USED TO TEACH READING CLASS
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                SOFTWARE             LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T046603              DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS                NUMBER OF SPECIFICATION RECORDS:  4

001   SOFTWA-1   (1,2         )  000         RDNG COMP SFTWRE (TEACHER): ALMOST EVERY DY,ONCE/TWICE WEEK
002   SOFTWA-2   (3           )  100         READING COMPUTER SOFTWARE (TEACHER):  ONCE OR TWICE A MONTH
003   SOFTWA-3   (4           )  010         READING COMPUTER SOFTWARE (TEACHER):  NEVER OF HARDLEY EVER
004   SOFTWA-?   (M,DNA       )  001         RDNG COMPUTER SOFTWARE (TEACHER):  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0012
DESCRIPTION:                HOW OFTEN ARE VARIETY OF BOOKS USED TO TEACH READING CLASS
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                VARI_BKS             LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T046604              DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS                NUMBER OF SPECIFICATION RECORDS:  4

001   VARBKS-1   (1           )  000         VARIETY OF BOOKS:  ALMOST EVERY DAY
002   VARBKS-2   (2           )  100         VARIETY OF BOOKS:  ONCE OR TWICE A WEEK
003   VARBKS-3   (3,4         )  010         VARIETY OF BOOKS:  ONCE OR TWICE A MONTH, NEVER/HARDLY EVER
004   VARBKS-?   (M,DNA       )  001         VARIETY OF BOOKS:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0013
DESCRIPTION:                HOW OFTEN ARE MATERIALS FROM OTHER SUBJECTS USED TO TEACH READING CLASS
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                OTH_MATS             LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T046605              DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS                NUMBER OF SPECIFICATION RECORDS:  4

001   OTHMAT-1   (1           )  000         OTHER SUBJECT MATERIALS (TEACHER):  ALMOST EVERY DAY
002   OTHMAT-2   (2           )  100         OTHER SUBJECT MATERIALS (TEACHER):  ONCE OR TWICE A WEEK
003   OTHMAT-3   (3,4         )  010         OTHER SUBJ MATRLS (TEACHER):  ONCE/TWICE MONTH, NEVER/HARDLY
004   OTHMAT-?   (M,DNA       )  001         OTHER SUBJECT MATERIALS (TEACHER):  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0014
DESCRIPTION:                HOW DO YOU DISCUSS NEW OR DIFFICULT VOCABULARY?
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                T_VOCAB              LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    T046701              DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS                NUMBER OF SPECIFICATION RECORDS:  5

001   VOCAB-T1   (1           )  0000        DISCUSS VOCABULARY:  ALMOST EVERY DAY
002   VOCAB-T2   (2           )  1000        DISCUSS VOCABULARY:  ONCE OR TWICE A WEEK
003   VOCAB-T3   (3           )  0100        DISCUSS VOCABULARY:  ONCE OR TWICE A MONTH
004   VOCAB-T4   (4           )  0010        DISCUSS VOCABULARY:  NEVER OR HARDLY EVER
005   VOCAB-T?   (M,DNA       )  0001        DISCUSS VOCABULARY:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0015
DESCRIPTION:                HOW OFTEN DO YOU ASK STUDENTS TO READ ALOUD?
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                T_ALOUD              LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    T046702              DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS                NUMBER OF SPECIFICATION RECORDS:  5

001   ALOUD-T1   (1           )  0000        READ ALOUD: ALMOST EVERY DAY
002   ALOUD-T2   (2           )  1000        READ ALOUD: ONCE OR TWICE A WEEK
003   ALOUD-T3   (3           )  0100        READ ALOUD: ONCE OR TWICE A MONTH
004   ALOUD-T4   (4           )  0010        READ ALOUD: NEVER OR HARDLY EVER
005   ALOUD-T?   (M,DNA       )  0001        READ ALOUD: MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0016
DESCRIPTION:                HOW OFTEN DO YOU ASK STUDENTS TO TALK TO EACH OTHER ABOUT WHAT THEY HAVE READ?
```

```
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               T_TALKRD           LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                   T046703            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:  5

001   TLKRD-T1   (1          )  0000          TALK ABOUT READING:  ALMOST EVERY DAY
002   TLKRD-T2   (2          )  1000          TALK ABOUT READING:  ONCE OR TWICE A WEEK
003   TLKRD-T3   (3          )  0100          TALK ABOUT READING:  ONCE OR TWICE A MONTH
004   TLKRD-T4   (4          )  0010          TALK ABOUT READING:  NEVER OR HARDLY EVER
005   TLKRD-T?   (M,DNA      )  0001          TALK ABOUT READING:  MISSING, DOES NOT APPLY

CONDITIONING ID:           TRED0017
DESCRIPTION:               HOW OFTEN DO YOU ASK STUDENTS TO WRITE SOMETHING ABOUT WHAT THEY HAVE READ?
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               T_WRITRD           LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                   T046704            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:  5

001   WRTRD-T1   (1          )  0000          WRITE ABOUT READING:  ALMOST EVERY DAY
002   WRTRD-T2   (2          )  1000          WRITE ABOUT READING:  ONCE OR TWICE A WEEK
003   WRTRD-T3   (3          )  0100          WRITE ABOUT READING:  ONCE OR TWICE A MONTH
004   WRTRD-T4   (4          )  0010          WRITE ABOUT READING:  NEVER OR HARDLY EVER
005   WRTRD-T?   (M,DNA      )  0001          WRITE ABOUT READING:  MISSING, DOES NOT APPLY

CONDITIONING ID:           TRED0018
DESCRIPTION:               HOW OFTEN DO YOU ASK STUDENTS TO WORK IN A READING WORKBOOK OR ON A WORKSHEET?
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               T_WBKWSH           LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                   T046705            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:  5

001   WB/WS-T1   (1          )  0000          READING WORKBOOK/WORKSHEET:  ALMOST EVERY DAY
002   WB/WS-T2   (2          )  1000          READING WORKBOOK/WORKSHEET:  ONCE OR TWICE A WEEK
003   WB/WS-T3   (3          )  0100          READING WORKBOOK/WORKSHEET:  ONCE OR TWICE A MONTH
004   WB/WS-T4   (4          )  0010          READING WORKBOOK/WORKSHEET:  NEVER OR HARDLY EVER
005   WB/WS-T?   (M,DNA      )  0001          READING WORKBOOK/WORKSHEET:  MISSING, MISSING NOT APPLY

CONDITIONING ID:           TRED0019
DESCRIPTION:               HOW OFTEN DO YOU ASK STUDENTS TO READ SILENTLY?
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               T_SILENT           LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                   T046706            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:  5

001   SILNT-T1   (1          )  0000          READ SILENTLY: ALMOST EVERY DAY
002   SILNT-T2   (2          )  1000          READ SILENTLY: ONCE OR TWICE A WEEK
003   SILNT-T3   (3          )  0100          READ SILENTLY: ONCE OR TWICE A MONTH
004   SILNT-T4   (4          )  0010          READ SILENTLY: NEVER OR HARDLY EVER
005   SILNT-T?   (M,DNA      )  0001          READ SILENTLY: MISSING, DOES NOT APPLY

CONDITIONING ID:           TRED0020
DESCRIPTION:               HOW OFTEN DO YOU GIVE STUDENTS TIME TO READ BOOKS OF THEIR OWN CHOOSING?
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               T_OWNBKS           LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                   T046707            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:  5

001   OWNBK-T1   (1          )  0000          BOOKS CHOSEN YOURSELF:  ALMOST EVERY DAY
002   OWNBK-T2   (2          )  1000          BOOKS CHOSEN YOURSELF:  OR TWICE A WEEK
003   OWNBK-T3   (3          )  0100          BOOKS CHOSEN YOURSELF:  ONCE OR TWICE A MONTH
004   OWNBK-T4   (4          )  0010          BOOKS CHOSEN YOURSELF:  NEVER OR HARDLY EVER
005   OWNBK-T?   (M,DNA      )  0001          BOOKS CHOSEN YOURSELF:  MISSING, DOES NOT APPLY

CONDITIONING ID:           TRED0021
DESCRIPTION:               HOW OFTEN YOU ASK STUDENTS TO WRITE IN LOG OR JOURNAL ABOUT WHAT THEY HAVE READ?
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               T_RDLOG            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                   T046708            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:  5
```

```
001   RDLOG-T1   (1          )  0000        WRITE IN LOG/JOURNAL:  ALMOST EVERY DAY
002   RDLOG-T2   (2          )  1000        WRITE IN LOG/JOURNAL:  ONCE OR TWICE A WEEK
003   RDLOG-T3   (3          )  0100        WRITE IN LOG/JOURNAL:  ONCE OR TWICE A MONTH
004   RDLOG-T4   (4          )  0010        WRITE IN LOG/JOURNAL:  NEVER OR HARDLY EVER
005   RDLOG-T?   (M,DNA      )  0001        WRITE IN LOG/JOURNAL:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0022
DESCRIPTION:                HOW OFTEN YOU ASK STUDENTS TO DO GROUP ACTIVITY/PROJECT ABOUT WHAT THEY READ?
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                T_RDPROJ            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                    T046709             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  5

001   RDPRJ-T1   (1          )  0000        PROJECT ABOUT READING:  ALMOST EVERY DAY
002   RDPRJ-T2   (2          )  1000        PROJECT ABOUT READING:  ONCE OR TWICE A WEEK
003   RDPRJ-T3   (3          )  0100        PROJECT ABOUT READING:  ONCE OR TWICE A MONTH
004   RDPRJ-T4   (4          )  0010        PROJECT ABOUT READING:  NEVER OR HARDLY EVER
005   RDPRJ-T?   (M,DNA      )  0001        PROJECT ABOUT READING:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0023
DESCRIPTION:                HOW MUCH READING INSTRUCTIONAL TIME DO YOU DEVOTE TO DECODING SKILLS?
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                DECODING            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T046801             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  4

001   DECODE-1   (1          )  000         DECODING SKILLS:  ALMOST ALL OF THE TIME
002   DECODE-2   (2          )  100         DECODING SKILLS:  SOME OF THE TIME
003   DECODE-3   (3          )  010         DECODING SKILLS:  NEVER OF HARDLY EVER
004   DECODE-?   (M,DNA      )  001         DECODING SKILLS:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0024
DESCRIPTION:                HOW MUCH READING INSTRUCTIONAL TIME DO YOU DEVOTE TO ORAL READING?
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                ORALREAD            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T046802             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  4

001   ORALRD-1   (1          )  000         ORAL READING:  ALMOST ALL OF THE TIME
002   ORALRD-2   (2          )  100         ORAL READING:  SOME OF THE TIME
003   ORALRD-3   (3          )  010         ORAL READING:  NEVER OF HARDLY EVER
004   ORALRD-?   (M,DNA      )  001         ORAL READING:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0025
DESCRIPTION:                HOW MUCH READING INSTRUCTIONAL TIME DO YOU DEVOTE TO VOCABULARY?
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                VOCABLRY            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T046803             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  4

001   VOCABY-1   (1          )  000         VOCABULARY:  ALMOST ALL OF THE TIME
002   VOCABY-2   (2          )  100         VOCABULARY:  SOME OF THE TIME
003   VOCABY-3   (3          )  010         VOCABULARY:  NEVER OF HARDLY EVER
004   VOCABY-?   (M,DNA      )  001         VOCABULARY:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0026
DESCRIPTION:                HOW MUCH RDNG INSTRUCT TIME DO YOU DEVOTE TO COMPREHENSION/INTERPRETATION?
GRADES/ASSESSMENTS:         N04, S04
GROUP LABEL:                COMPREH             LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                    T046804             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:           CLASS               NUMBER OF SPECIFICATION RECORDS:  4

001   CMPREH-1   (1          )  000         COMPREHENSION/INTERPRETATION:  ALMOST ALL OF THE TIME
002   CMPREH-2   (2          )  100         COMPREHENSION/INTERPRETATION:  SOME OF THE TIME
003   CMPREH-3   (3          )  010         COMPREHENSION/INTERPRETATION:  NEVER OF HARDLY EVER
004   CMPREH-?   (M,DNA      )  001         COMPREHENSION/INTERPRETATION:  MISSING, DOES NOT APPLY

CONDITIONING ID:            TRED0027
DESCRIPTION:                HOW MUCH READING INSTRUCTIONAL TIME DO YOU DEVOTE TO READING STRATEGIES?
```

```
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               STRATGY            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                   T046805            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:  4

001   STRATG-1   (1          )  000         READING STRATEGIES:  ALMOST ALL OF THE TIME
002   STRATG-2   (2          )  100         READING STRATEGIES:  SOME OF THE TIME
003   STRATG-3   (3          )  010         READING STRATEGIES:  NEVER OF HARDLY EVER
004   STRATG-?   (M,DNA      )  001         READING STRATEGIES:  MISSING, DOES NOT APPLY

CONDITIONING ID:           TRED0028
DESCRIPTION:               HOW MUCH EMPHASIS DO YOU GIVE PHONICS?
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               EMP_PHON           LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                   T046901            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:  4

001   EMPPHO-H   (1          )  000         EMPHASIS PHONICS:  HEAVY
002   EMPPHO-M   (2          )  100         EMPHASIS PHONICS:  MODERATE
003   EMPPHO-L   (3          )  010         EMPHASIS PHONICS:  LITTLE OR NO
004   EMPPHO-?   (M,DNA      )  001         EMPHASIS PHONICS:  MISSING, DOES NOT APPLY

CONDITIONING ID:           TRED0029
DESCRIPTION:               HOW MUCH EMPHASIS DO YOU GIVE THE INTEGRATION OF READING AND WRITING?
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               EMP_INTG           LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                   T046902            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:  4

001   EMPINT-H   (1          )  000         EMPHASIS INTEGRATION READING/WRITING:  HEAVY
002   EMPINT-M   (2          )  100         EMPHASIS INTEGRATION READING/WRITING:  MODERATE
003   EMPINT-L   (3          )  010         EMPHASIS INTEGRATION READING/WRITING:  LITTLE OR NO
004   EMPINT-?   (M,DNA      )  001         EMPHASIS INTEGRATION READING/WRITING:  MISSING, DOESNT APPLY

CONDITIONING ID:           TRED0030
DESCRIPTION:               HOW MUCH EMPHASIS DO YOU GIVE WHOLE LANGUAGE?
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               EMP_WLAN           LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                   T046903            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:  4

001   EMPLAN-H   (1          )  000         EMPHASIS WHOLE LANGUAGE:  HEAVY
002   EMPLAN-M   (2          )  100         EMPHASIS WHOLE LANGUAGE:  MODERATE
003   EMPLAN-L   (3          )  010         EMPHASIS WHOLE LANGUAGE:  LITTLE OR NO
004   EMPLAN-?   (M,DNA      )  001         EMPHASIS WHOLE LANGUAGE:  MISSING, DOES NOT APPLY

CONDITIONING ID:           TRED0031
DESCRIPTION:               HOW MUCH EMPHASIS DO YOU GIVE LITERATURE-BASED READING?
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               EMP_LITB           LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                   T046904            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:  4

001   EMPLIT-H   (1          )  000         EMPHASIS LITERATURE-BASED READING:  HEAVY
002   EMPLIT-M   (2          )  100         EMPHASIS LITERATURE-BASED READING:  MODERATE
003   EMPLIT-L   (3          )  010         EMPHASIS LITERATURE-BASED READING:  LITTLE OR NO
004   EMPLIT-?   (M,DNA      )  001         EMPHASIS LITERATURE-BASED READING:  MISSING, DOES NOT APPLY

CONDITIONING ID:           TRED0032
DESCRIPTION:               HOW MUCH EMPHASIS DO YOU GIVE READING ACROSS THE CONTENT AREAS?
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               EMP_CONT           LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                   T046905            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS              NUMBER OF SPECIFICATION RECORDS:  4

001   EMPCON-H   (1          )  000         EMPHASIS READING ACROSS CONTENT AREAS:  HEAVY
002   EMPCON-M   (2          )  100         EMPHASIS READING ACROSS CONTENT AREAS:  MODERATE
003   EMPCON-L   (3          )  010         EMPHASIS READING ACROSS CONTENT AREAS:  LITTLE OR NO
004   EMPCON-?   (M,DNA      )  001         EMPHASIS READING ACROSS CONTENT AREAS:  MSSNG, DOESNT APPLY
```

```
CONDITIONING ID:          TRED0033
DESCRIPTION:              HOW MUCH EMPHASIS DO YOU GIVE INDIVIDUALIZED READING PROGRAMS?
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              EMP_INDV            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                  T046906             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  4

001   EMPCON-H   (1          )  000           EMPHASIS INDIVIDUALIZED READING PROGRAMS:  HEAVY
002   EMPCON-M   (2          )  100           EMPHASIS INDIVIDUALIZED READING PROGRAMS:  MODERATE
003   EMPCON-L   (3          )  010           EMPHASIS INDIVIDUALIZED READING PROGRAMS:  LITTLE OR NO
004   EMPCON-?   (M,DNA      )  001           EMPHASIS INDIVIDUALIZED RDING PROGRAMS: MSSNG, DOESNT APPLY

CONDITIONING ID:          TRED0034
DESCRIPTION:              HOW OFTEN DO YOU USE MULTIPLE-CHOICE TESTS TO ASSESS STUDENTS IN READING?
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              MC_TESTS            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                  T047001             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  4

001   MCTEST-1   (1          )  000           MULTIPLE-CHOICE TESTS:  ONCE OR TWICE A WEEK
002   MCTEST-2   (2          )  100           MULTIPLE-CHOICE TESTS:  ONCE OR TWICE A MONTH
003   MCTEST-3   (3,4        )  010           MULTIPLE-CHOICE TESTS:  ONCE/TWICE YEAR, NEVER/HARDLY EVER
004   MCTEST-?   (M,DNA      )  001           MULTIPLE-CHOICE TESTS:  MISSING, DOES NOT APPLY

CONDITIONING ID:          TRED0035
DESCRIPTION:              HOW OFTEN DO YOU USE MULTIPLE-CHOICE TESTS TO ASSESS STUDENTS IN READING?
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              SA_TESTS            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                  T047002             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  4

001   SATEST-1   (1          )  000           SHORT-ANSWER TESTS:  ONCE OR TWICE A WEEK
002   SATEST-2   (2          )  100           SHORT-ANSWER TESTS:  ONCE OR TWICE A MONTH
003   SATEST-3   (3,4        )  010           SHORT-ANSWER TESTS:  ONCE OR TWICE A YEAR, NEVER/HARDLY EVER
004   SATEST-?   (M,DNA      )  001           SHORT-ANSWER TESTS:  MISSING, DOES NOT APPLY

CONDITIONING ID:          TRED0036
DESCRIPTION:              HOW OFTEN DO YOU USE WRITING PARAGRAPHS TO ASSESS STUDENTS IN READING?
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              WRT_PARA            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                  T047003             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  4

001   WRTPAR-1   (1          )  000           ASSESS BY WRITING PARAGRAPHS:  ONCE OR TWICE A WEEK
002   WRTPAR-2   (2          )  100           ASSESS BY WRITING PARAGRAPHS:  ONCE OR TWICE A MONTH
003   WRTPAR-3   (3,4        )  010           ASSESS BY WRITING PARAGRAPHS:  ONCE/TWICE YEAR, NEVER/HARDLY
004   WRTPAR-?   (M,DNA      )  001           ASSESS BY WRITING PARAGRAPHS:  MISSING, DOES NOT APPLY

CONDITIONING ID:          TRED0037
DESCRIPTION:              HOW OFTEN DO YOU USE OBSERVATIONS TO ASSESS STUDENTS IN READING?
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              OBSERVTN            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                  T047004             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  4

001   OBSERV-1   (1          )  000           ASSESS BY OBSERVATIONS:  ONCE OR TWICE A WEEK
002   OBSERV-2   (2          )  100           ASSESS BY OBSERVATIONS:  ONCE OR TWICE A MONTH
003   OBSERV-3   (3,4        )  010           ASSESS BY OBSERVATIONS:  ONCE/TWICE YEAR, NEVER/HARDLY EVER
004   OBSERV-?   (M,DNA      )  001           ASSESS BY OBSERVATIONS:  MISSING, DOES NOT APPLY

CONDITIONING ID:          TRED0038
DESCRIPTION:              HOW OFTEN DO YOU USE ORAL READING TO ASSESS STUDENTS IN READING?
GRADES/ASSESSMENTS:       N04, S04
GROUP LABEL:              ORALTEST            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                  T047005             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:         CLASS               NUMBER OF SPECIFICATION RECORDS:  4

001   ORLTST-1   (1          )  000           ASSESS BY ORAL READING:  ONCE OR TWICE A WEEK
```

```
002  ORLTST-2  (2          )  100        ASSESS BY ORAL READING:  ONCE OR TWICE A MONTH
003  ORLTST-3  (3,4        )  010        ASSESS BY ORAL READING:  ONCE/TWICE YEAR, NEVER/HARDLY EVER
004  ORLTST-?  (M,DNA      )  001        ASSESS BY ORAL READING:  MISSING, DOES NOT APPLY

CONDITIONING ID:             TRED0039
DESCRIPTION:                 HOW OFTEN USE INDIVID OR GROUP PROJECTS/PRESENTTNS TO ASSESS STUDENTS IN RDING?
GRADES/ASSESSMENTS:          N04, S04
GROUP LABEL:                 PROJECTS            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                     T047006             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:            CLASS               NUMBER OF SPECIFICATION RECORDS:  4

001  PROJCT-1  (1          )  000        ASSESS BY PROJECTS/PRESENTATIONS:  ONCE OR TWICE A WEEK
002  PROJCT-2  (2          )  100        ASSESS BY PROJECTS/PRESENTATIONS:  ONCE OR TWICE A MONTH
003  PROJCT-3  (3,4        )  010        ASSESS BY PROJECTS/PRESENTTNS:ONCE/TWICE/YEAR, NEVER/HARDLY
004  PROJCT-?  (M,DNA      )  001        ASSESS BY PROJECTS/PRESENTATIONS:  MISSING, DOES NOT APPLY

CONDITIONING ID:             TRED0040
DESCRIPTION:                 HOW OFTEN DO YOU USE READING PORTFOLIOS TO ASSESS STUDENTS IN READING?
GRADES/ASSESSMENTS:          N04, S04
GROUP LABEL:                 RD_PORTF            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                     T047007             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:            CLASS               NUMBER OF SPECIFICATION RECORDS:  4

001  RDPORT-1  (1          )  000        ASSESS BY READING PORTFOLIOS:  ONCE OR TWICE A WEEK
002  RDPORT-2  (2          )  100        ASSESS BY READING PORTFOLIOS:  ONCE OR TWICE A MONTH
003  RDPORT-3  (3,4        )  010        ASSESS BY READING PORTFOLIOS:  ONCE/TWICE/YEAR, NEVER/HARDLY
004  RDPORT-?  (M,DNA      )  001        ASSESS BY READING PORTFOLIOS:  MISSING, DOES NOT APPLY

CONDITIONING ID:             TRED0041
DESCRIPTION:                 HOW OFTEN DO YOU USE STUDENT SELF-REPORTS TO ASSESS STUDENTS IN READING?
GRADES/ASSESSMENTS:          N04, S04
GROUP LABEL:                 SELF_REP            LENGTH OF CONTRAST FIELD       :  3
NAEP ID:                     T047008             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:            CLASS               NUMBER OF SPECIFICATION RECORDS:  4

001  SLFREP-1  (1          )  000        ASSESS BY STUDENT SELF-REPORTS:  ONCE OR TWICE A WEEK
002  SLFREP-2  (2          )  100        ASSESS BY STUDENT SELF-REPORTS:  ONCE OR TWICE A MONTH
003  SLFREP-3  (3,4        )  010        ASSESS BY STUDENT SELF-REPORTS:  ONCE/TWICE/YR, NEVER/HARDLY
004  SLFREP-?  (M,DNA      )  001        ASSESS BY STUDENT SELF-REPORTS:  MISSING, DOES NOT APPLY

CONDITIONING ID:             TRED0042
DESCRIPTION:                 HOW OFTEN DO YOU SEND OR TAKE THE CLASS TO THE LIBRARY?
GRADES/ASSESSMENTS:          N04, S04
GROUP LABEL:                 CLA2LIBR            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                     T047101             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:            CLASS               NUMBER OF SPECIFICATION RECORDS:  5

001  CLALIB-1  (1          )  0000       TAKE CLASS TO LIBRARY:  ALMOST EVERY DAY
002  CLALIB-2  (2          )  1000       TAKE CLASS TO LIBRARY:  ONCE OR TWICE A WEEK
003  CLALIB-3  (3          )  0100       TAKE CLASS TO LIBRARY:  ONCE OR TWICE A MONTH
004  CLALIB-4  (4          )  0010       TAKE CLASS TO LIBRARY:  NEVER OR HARDLY EVER
005  CLALIB-5  (5,M,DNA    )  0001       TAKE CLASS TO LIBRARY:  NO LIBRARY, MISSING, DOES NOT APPLY

CONDITIONING ID:             TRED0043
DESCRIPTION:                 HOW OFTEN DO YOU ASSIGN STUDENTS TO READ A BOOK FROM THE LIBRARY?
GRADES/ASSESSMENTS:          N04, S04
GROUP LABEL:                 LIB_BOOK            LENGTH OF CONTRAST FIELD       :  4
NAEP ID:                     T047102             DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:            CLASS               NUMBER OF SPECIFICATION RECORDS:  5

001  CLALIB-1  (1          )  0000       ASSIGN STUDENTS TO READ LIBRARY BOOK:  ALMOST EVERY DAY
002  CLALIB-2  (2          )  1000       ASSIGN STUDENTS TO READ LIBRARY BOOK:  ONCE OR TWICE A WEEK
003  CLALIB-3  (3          )  0100       ASSIGN STUDENTS TO READ LIBRARY BOOK:  ONCE OR TWICE A MONTH
004  CLALIB-4  (4          )  0010       ASSIGN STUDENTS TO READ LIBRARY BOOK:  NEVER OR HARDLY EVER
005  CLALIB-5  (5,M,DNA    )  0001       ASSIGN STUDENTS TO READ LIBRARY BOOK: NO LIB,MSSNG,NOT APPLY

CONDITIONING ID:             TRED0044
DESCRIPTION:                 ARE COMPUTERS AVAILABLE FOR USE BY STUDENTS IN READING CLASS?
GRADES/ASSESSMENTS:          N04, S04
```

```
GROUP LABEL:               COMP4RDG          LENGTH OF CONTRAST FIELD      :  3
NAEP ID:                   T047201           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS             NUMBER OF SPECIFICATION RECORDS:  4

001   COMP-NA    (1          )  000          COMPUTERS IN READING CLASS:  NOT AVAILABLE
002   COMP-DIF   (2          )  100          COMPUTERS IN READING CLASS:  AVAILABLE BUT DIFF TO ACCESS
003   COMP-AVL   (3          )  010          COMPUTERS IN READING CLASS:  AVAILABLE IN THE CLASSROOM
004   COMP-?     (M,DNA      )  001          COMPUTERS IN READING CLASS:  MISSING, DOES NOT APPLY

CONDITIONING ID:           TRED0045
DESCRIPTION:               OTHER TEACHING CERTIFICATION (DERIVED)
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               CERTOTH           LENGTH OF CONTRAST FIELD      :  3
NAEP ID:                   TRCERT            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS             NUMBER OF SPECIFICATION RECORDS:  4

001   MT/OTH-Y   (1          )  000         MIDDLE/JUNIOR HIGH/SECNDRY MATH OR OTHER: YES
002   MATH-NO    (2          )  100         MIDDLE/JUNIOR HIGH/SECNDRY MATH: NO
003   MATH-NS    (3          )  010         MIDDLE/JUNIOR HIGH/SECNDRY MATH: NOT OFFERED IN STATE
004   M/OTH-?    (M,DNA      )  001         MIDDLE/JR HIGH/SECNDRY MATH: MSSNG; OTHER: NO, NOT OFFERE

CONDITIONING ID:           TRED0046
DESCRIPTION:               OTHER UNDERGRADUATE MAJOR (DERIVED)
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               TRUMAJB           LENGTH OF CONTRAST FIELD      :  2
NAEP ID:                   TRUMAJB           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS             NUMBER OF SPECIFICATION RECORDS:  3

001   UMAJO-Y    (1          )  00           OTHER OR (MATHEMATICS AND MATHEMATICS EDUCATION): YES
002   UMAJO-N    (2          )  10           OTHER AND (MATHEMATICS OR MATHEMATICS EDUCATION): MISSING
003   UMAJO-?    (M,DNA      )  01           OTHER AND (MATHEMATICS OR MATHEMATICS EDUCATION): MISSING

CONDITIONING ID:           TRED0047
DESCRIPTION:               OTHER GRADUATE MAJOR (DERIVED)
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               TRGMAJB           LENGTH OF CONTRAST FIELD      :  2
NAEP ID:                   TRGMAJB           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS             NUMBER OF SPECIFICATION RECORDS:  3

001   GMAJO-Y    (1          )  00           OTHER OR (MATHEMATICS AND MATHEMATICS EDUCATION): YES
002   GMAJO-N    (2          )  10           OTHER AND (MATHEMATICS OR MATHEMATICS EDUCATION): MISSING
003   GMAJO-?    (M,DNA      )  01           OTHER AND (MATHEMATICS OR MATHEMATICS EDUCATION): MISSING

CONDITIONING ID:           TRED0048
DESCRIPTION:               SUM OF 'YES' RESPONSES TO TEACHER READING TRAINING VARIABLES (DERIVED)
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               TRTRAIN           LENGTH OF CONTRAST FIELD      :  4
NAEP ID:                   TRTRAIN           DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS             NUMBER OF SPECIFICATION RECORDS:  5

001   TRNS0-3    (1          )  0000        SUM(YES) = 0-3
002   TRNS4-6    (2          )  1000        SUM(YES) = 4-6
003   TRNS7      (3          )  0100        SUM(YES) = 7
004   TRNS8      (4          )  0010        SUM(YES) = 8
005   TRNS?      (M,DNA      )  0001        SUM(YES) = ?

CONDITIONING ID:           TRED0049
DESCRIPTION:               TEACHER EMPHASIS VARIABLE 1 (DERIVED)
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               TREMP1            LENGTH OF CONTRAST FIELD      :  2
NAEP ID:                   TREMP1            DEGREES OF FREEDOM PER CONTRAST:  1
TYPE OF CONTRAST:          CLASS             NUMBER OF SPECIFICATION RECORDS:  3

001   TREMP1-1   (1          )  00           DECODING SKILLS AND PHONICS: BOTH HEAVY EMPHASIS
002   TREMP1-2   (2          )  10           DECODING SKILLS OR PHONICS: OTHERWISE
003   TREMP1-?   (DNA        )  01           DECODING SKILLS OR PHONICS: OTHERWISE

CONDITIONING ID:           TRED0050
DESCRIPTION:               TEACHER EMPHASIS VARIABLE 2 (DERIVED)
```

BEST COPY AVAILABLE

```
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               TREMP2              LENGTH OF CONTRAST FIELD       :   2
NAEP ID:                   TREMP2              DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:          CLASS               NUMBER OF SPECIFICATION RECORDS:   3

001   TREMP2-1   (1            )  00           VOCAB/INTEGRATION OF RDNG AND WRTNG/WHOLE LANG: 2 OR MORE
002   TREMP2-2   (2            )  10           VOCAB/INTEGRATION OF RDNG AND WRTNG/WHOLE LANG: OTHERWISE
003   TREMP2-7   (DNA          )  01           VOCAB/INTEGRATION OF RDNG AND WRTNG/WHOLE LANG: OTHERWISE

CONDITIONING ID:           TRED0051
DESCRIPTION:               TEACHER EMPHASIS VARIABLE 3 (DERIVED)
GRADES/ASSESSMENTS:        N04, S04
GROUP LABEL:               TREMP3              LENGTH OF CONTRAST FIELD       :   2
NAEP ID:                   TREMP3              DEGREES OF FREEDOM PER CONTRAST:   1
TYPE OF CONTRAST:          CLASS               NUMBER OF SPECIFICATION RECORDS:   3

001   TREMP3-1   (1            )  00           RDNG ACROSS CONTENT AREAS & INDIV READING PROGRAMS:BOTH HEAV
002   TREMP3-2   (2            )  10           RDNG ACROSS CONTENT AREAS & INDIV READING PROGRAMS:OTHERWISE
003   TREMP3-7   (DNA          )  01           RDNG ACROSS CONTENT AREAS & INDIV READING PROGRAMS:OTHERWISE
```

APPENDIX D

IRT PARAMETERS FOR READING ITEMS

APPENDIX D

IRT Parameters

This appendix contains 2 tables of IRT (item response theory) parameters for the items that were used in each reading scale for the fourth-grade Trial State Assessment.

For each of the binary scored items used in scaling (i.e., multiple-choice items and short constructed-response items), the tables provide estimates of the IRT parameters (which correspond to a_j, b_j, and c_j in equation 8.1 in Chapter 8) and their associated standard errors (s.e.) of the estimates. For each of the polytomously scored items (i.e., the extended constructed-response items), the tables also show the estimates of the d_{jv} parameters (see equation 8.1) and their associated standard errors.

The tables also show the block in which each item appears (*Block*) and the position of each item within its block (*Item*).

Note that the item parameters in this appendix are in the metrics used for the original calibration of the scales. The transformations needed to represent these parameters in terms of the metrics of the final reporting scales are given in Chapter 9.

Table D-1
IRT Parameters for Grade 4 Reading Items
Reading for Literary Experience

NAEP ID	*Block*	*Item*	a_j *(s.e.)*	b_j *(s.e.)*	c_j *(s.e.)*	d_{j1} *(s.e.)*	d_{j2} *(s.e.)*	d_{j3} *(s.e.)*
R012111	RD	11	0.955 (0.025)	1.598 (0.020)	0.000 (0.000)	1.206 (0.020)	-1.206 (0.065)	
R012107	RD	7	1.444 (0.072)	0.273 (0.030)	0.285 (0.014)			
R012004	RC	4	0.783 (0.024)	0.399 (0.021)	0.000 (0.000)			
R012409	RI	9	1.106 (0.036)	0.606 (0.019)	0.000 (0.000)			
R012611	RE	11	0.703 (0.024)	-0.513 (0.032)	0.000 (0.000)			
R012607	RE	7	1.050 (0.028)	1.964 (0.015)	0.000 (0.000)	1.430 (0.020)	0.753 (0.026)	-2.183 (0.286)
R012101	RD	1	1.443 (0.071)	-1.117 (0.053)	0.347 (0.026)			
R012106	RD	6	0.810 (0.025)	0.345 (0.021)	0.000 (0.000)			
R012110	RD	10	0.763 (0.042)	-1.267 (0.116)	0.266 (0.042)			
R012011	RC	11	1.554 (0.072)	-0.283 (0.035)	0.253 (0.018)			
R012007	RC	7	0.744 (0.041)	-0.601 (0.087)	0.225 (0.031)			
R012403	RI	3	0.917 (0.029)	0.816 (0.023)	0.000 (0.000)			
R012601	RE	1	0.877 (0.031)	1.199 (0.032)	0.000 (0.000)			
R012408	RI	8	1.391 (0.076)	0.187 (0.037)	0.307 (0.017)			
R012610	RE	10	1.779 (0.116)	0.667 (0.027)	0.353 (0.012)			
R012606	RE	6	1.810 (0.097)	0.493 (0.024)	0.322 (0.012)			
R012001	RC	1	1.269 (0.056)	0.606 (0.022)	0.106 (0.010)			
R012109	RD	9	0.515 (0.020)	-1.309 (0.056)	0.000 (0.000)			
R012010	RC	10	1.130 (0.033)	-0.720 (0.025)	0.000 (0.000)			
R012006	RC	6	0.481 (0.014)	0.772 (0.021)	0.000 (0.000)	0.669 (0.041)	-0.003 (0.046)	-0.667 (0.059)
R012402	RI	2	1.143 (0.074)	0.123 (0.055)	0.449 (0.019)			
R012609	RE	9	0.941 (0.077)	0.899 (0.045)	0.259 (0.017)			
R012103	RD	3	1.084 (0.044)	-0.637 (0.045)	0.178 (0.020)			
R012112	RD	12	0.713 (0.026)	-0.673 (0.037)	0.000 (0.000)			
R012108	RD	8	0.635 (0.021)	-1.323 (0.046)	0.000 (0.000)			
R012009	RC	9	1.128 (0.060)	-0.835 (0.071)	0.340 (0.030)			
R012405	RI	5	1.239 (0.072)	0.761 (0.028)	0.196 (0.012)			
R012603	RE	3	1.670 (0.074)	0.182 (0.025)	0.250 (0.013)			
R012608	RE	8	0.626 (0.044)	-0.693 (0.140)	0.293 (0.042)			
R012102	RD	2	0.786 (0.023)	-1.352 (0.037)	0.000 (0.000)			
R012003	RC	3	1.776 (0.067)	-0.582 (0.026)	0.178 (0.015)			
R012008	RC	8	0.949 (0.028)	-0.726 (0.027)	0.000 (0.000)			
R012404	RI	4	0.918 (0.052)	0.161 (0.051)	0.242 (0.020)			
R012602	RE	2	1.440 (0.082)	1.321 (0.031)	0.168 (0.007)			
R012105	RD	5	0.733 (0.042)	-0.089 (0.070)	0.201 (0.025)			
R012002	RC	2	1.360 (0.033)	-0.134 (0.015)	0.000 (0.000)			
R012407	RI	7	0.771 (0.024)	-0.652 (0.030)	0.000 (0.000)			
R012605	RE	5	1.021 (0.084)	0.993 (0.041)	0.307 (0.014)			
R012104	RD	4	0.629 (0.020)	-0.394 (0.029)	0.000 (0.000)			
R012005	RC	5	1.004 (0.048)	0.066 (0.041)	0.202 (0.017)			
R012401	RI	1	0.867 (0.020)	1.740 (0.019)	0.000 (0.000)	1.374 (0.020)	-0.708 (0.048)	-0.666 (0.123)
R012406	RI	6	0.792 (0.025)	0.325 (0.022)	0.000 (0.000)			
R012604	RE	4	0.992 (0.035)	1.212 (0.030)	0.000 (0.000)			

Table D-2
IRT Parameters for Grade 4 Reading Items
Reading to Gain Information

NAEP ID	*Block*	*Item*	a_j *(s.e.)*	b_j *(s.e.)*	c_j *(s.e.)*	d_{j1} *(s.e.)*	d_{j2} *(s.e.)*	d_{j3} *(s.e.)*
R012305	RH	5	0.387 (0.007)	1.457 (0.029)	0.000 (0.000)	3.715 (0.055)	0.148 (0.044)	-3.862 (0.167)
R012202	RF	2	0.817 (0.055)	0.576 (0.054)	0.267 (0.019)			
R012503	RJ	3	0.798 (0.024)	0.654 (0.024)	0.000 (0.000)			
R012701	RG	1	1.259 (0.058)	-0.026 (0.037)	0.274 (0.017)			
R012512	RJ	12	0.429 (0.013)	0.841 (0.025)	0.000 (0.000)	1.147 (0.048)	0.056 (0.051)	-1.204 (0.074)
R012508	RJ	8	0.872 (0.025)	-0.329 (0.022)	0.000 (0.000)			
R012710	RG	10	0.972 (0.032)	0.597 (0.023)	0.000 (0.000)			
R012706	RG	6	0.562 (0.022)	1.070 (0.041)	0.000 (0.000)			
R012304	RH	4	1.931 (0.157)	2.256 (0.083)	0.255 (0.006)			
R012205	RF	5	1.315 (0.073)	0.562 (0.032)	0.273 (0.013)			
R012502	RJ	2	0.871 (0.045)	-1.920 (0.118)	0.284 (0.051)			
R012709	RG	9	0.728 (0.061)	0.305 (0.096)	0.346 (0.029)			
R012307	RH	7	1.256 (0.056)	-0.064 (0.036)	0.223 (0.017)			
R012204	RF	4	0.419 (0.011)	0.161 (0.020)	0.000 (0.000)	1.434 (0.050)	-0.344 (0.045)	-1.089 (0.055)
R012505	RJ	5	1.168 (0.057)	-0.614 (0.056)	0.324 (0.025)			
R012703	RG	3	1.092 (0.030)	0.739 (0.019)	0.000 (0.000)			
R012708	RG	8	0.708 (0.018)	2.082 (0.024)	0.000 (0.000)	1.809 (0.026)	-0.334 (0.051)	-1.475 (0.190)
R012301	RH	1	1.045 (0.063)	0.134 (0.055)	0.399 (0.019)			
R012310	RH	10	0.938 (0.029)	0.459 (0.021)	0.000 (0.000)			
R012306	RH	6	0.776 (0.024)	0.684 (0.024)	0.000 (0.000)			
R012207	RF	7	0.515 (0.036)	-0.650 (0.155)	0.248 (0.043)			
R012504	RJ	4	0.644 (0.020)	-0.344 (0.027)	0.000 (0.000)			
R012702	RG	2	0.607 (0.020)	-1.271 (0.044)	0.000 (0.000)			
R012201	RF	1	0.251 (0.015)	-0.211 (0.058)	0.000 (0.000)			
R012511	RJ	11	0.941 (0.028)	-0.613 (0.025)	0.000 (0.000)			
R012309	RH	9	0.670 (0.057)	0.847 (0.072)	0.248 (0.023)			
R012210	RF	10	0.603 (0.024)	-1.451 (0.056)	0.000 (0.000)			
R012206	RF	6	1.086 (0.032)	0.809 (0.021)	0.000 (0.000)			
R012507	RJ	7	1.377 (0.069)	-0.355 (0.045)	0.367 (0.021)			
R012705	RG	5	1.173 (0.048)	1.761 (0.041)	0.000 (0.000)			
R012303	RH	3	0.941 (0.025)	-0.393 (0.020)	0.000 (0.000)			
R012501	RJ	1	0.654 (0.158)	2.910 (0.348)	0.308 (0.013)			
R012510	RJ	10	0.996 (0.057)	-0.150 (0.062)	0.344 (0.024)			
R012308	RH	8	0.732 (0.023)	0.406 (0.024)	0.000 (0.000)			
R012209	RF	9	1.210 (0.060)	0.269 (0.034)	0.182 (0.015)			
R012506	RJ	6	0.733 (0.022)	-0.210 (0.023)	0.000 (0.000)			
R012704	RG	4	1.401 (0.065)	0.743 (0.023)	0.149 (0.009)			
R012302	RH	2	0.902 (0.041)	-0.373 (0.055)	0.206 (0.024)			
R012203	RF	3	0.722 (0.049)	0.678 (0.056)	0.204 (0.019)			
R012208	RF	8	0.740 (0.024)	-0.558 (0.029)	0.000 (0.000)			
R012509	RJ	9	0.639 (0.040)	-0.739 (0.130)	0.287 (0.041)			
R012707	RG	7	2.089 (0.099)	0.421 (0.020)	0.249 (0.011)			

APPENDIX E

TRIAL STATE ASSESSMENT REPORTING SUBGROUPS

COMPOSITE AND DERIVED COMMON BACKGROUND VARIABLES

COMPOSITE AND DERIVED REPORTING VARIABLES

REPORTING SUBGROUPS FOR THE 1992 TRIAL STATE ASSESSMENT

Results for the 1992 Trial State Assessment were reported for student subgroups defined by gender, race/ethnicity, type of community, parents' level of education, and geographical region. The following explains how each of these subgroups was derived.

DSEX (Gender)

The variable SEX is the gender of the student being assessed, as taken from school records. For a few students, data for this variable was missing and was imputed by ETS after the assessment. The resulting variable DSEX contains a value for every student and is used for gender comparisons among students.

DRACE (Race/ethnicity)

The variable DRACE is an imputed definition of race/ethnicity, derived from up to three sources of information. This variable is used for race/ethnicity subgroup comparisons. Two items from the student demographics questionnaire were used in the determination of derived race/ethnicity:

Demographic Item Number 2:

2. If you are Hispanic, what is your Hispanic background?

○ I am not Hispanic.
○ Mexican, Mexican American, or Chicano
○ Puerto Rican
○ Cuban
○ Other Spanish or Hispanic background

Students who responded to item number 2 by filling in the second, third, fourth, or fifth oval were considered Hispanic. For students who filled in the first oval, did not respond to the item, or provided information that was illegible or could not be classified, responses to item number 1 were examined in an effort to determine race/ethnicity. Item number 1 read as follows:

Demographic Item Number 1:

1. Which best describes you?

○ White (not Hispanic)

○ Black (not Hispanic)

○ Hispanic ("Hispanic" means someone who is Mexican, Mexican American, Chicano, Puerto Rican, Cuban, or from some other Spanish or Hispanic background.)

○ Asian or Pacific Islander ("Asian or Pacific Islander" means someone who is Chinese, Japanese, Korean, Filipino, Vietnamese, or from some other Asian or Pacific Island background.)

○ American Indian or Alaskan Native ("American Indian or Alaskan Native" means someone who is from one of the American Indian tribes, or one of the original people of Alaska.)

○ Other (What?) ________

Students' race/ethnicity was then assigned to correspond with their selection. For students who filled in the sixth oval ("Other"), provided illegible information or information that could not be classified, or did not respond at all, race/ethnicity as provided from school records was used.

Derived race/ethnicity could not be determined for students who did not respond to background items 1 or 2 and for whom race/ethnicity was not provided by the school.

TOC (Type of community)

NAEP assigned each participating school to one of four type of categories designed to provide information about the communities in which the schools are located.

The type of community categories consist of three "extreme" types of communities and one "other" type of community. Schools were placed into these categories on the basis of information about the type of community, the size of its population (as of the 1980 Census), and an occupational profile of residents provided by school principals before the assessment. The principals completed estimates of the percentage of students whose parents fit into each of six occupational categories. For those schools where the principal or his or her designate was unable or unwilling to answer the question on the occupational profile of parents, the type of community category was assigned as "missing."

The definitions of these "extreme" categories were determined using data from the 1992 national assessment. The categories are formed so that, approximately, an estimated 10 percent of the student population nationally at each grade level attend schools in each of the three "extreme" community types. These same criteria were then applied on a school-by-school basis to the schools that participated in the state assessments, to determine the type of community classification for each. The procedure for establishing these "extreme" classes using the national data has been similar throughout NAEP's history. This procedure is described in the sampling and weighting reports for the 1986, 1988, and 1990 national assessments (see Burke, Braden, Hansen, Lago, & Tepping, 1987; Rust, Bethel, Burke, & Hansen, 1990; and Rust, Burke, Fahimi, & Wallace, 1992). The type of community categories are as follows:

> 1 - Extreme Rural: Students in this group live outside metropolitan statistical areas, live in areas with a population below 10,000, and attend schools where many of the students' parents are farmers or farm workers.
>
> 2 - Disadvantaged Urban: Students in this group live in metropolitan statistical areas and attend schools where a high proportion of the students' parents are on welfare or are not regularly employed.
>
> 3 - Advantaged Urban: Students in this group live in metropolitan statistical areas and attend schools where a high proportion of the students' parents are in professional or managerial positions.
>
> 4 - Other: Students in this category attend schools in areas other than those defined as advantaged urban, disadvantaged urban, or extreme rural.

PARED (Parents' education level)

The variable PARED is derived from responses to two questions, B003501 and B003601, in the student demographic questionnaire. Students were asked to indicate the extent of their mother's education (B003501—How far in high school did your mother go?) by choosing one of the following:

○ She did not finish high school.
○ She graduated from high school.
○ She had some education after high school.
○ She graduated from college.
○ I don't know.

Students were asked to provide the same information about the extent of their father's education (B003601—How far in high school did your father go?) by choosing one of the following:

○ He did not finish high school.
○ He graduated from high school.
○ He had some education after high school.
○ He graduated from college.
○ I don't know.

The information was combined into one parental education reporting category (PARED) as follows: If a student indicated the extent of education for only one parent, that level was included in the data. If a student indicated the extent of education for both parents, the higher of the two levels was included in the data. For students who did not know the level of education for both parents or did not know the level of education for one parent and did not respond for the other, the parental education level was classified as unknown. If the student did not respond for both parents, the student was recorded as having provided no response.

REGION (Region of the country)

States were grouped into four geographical regions—Northeast, Southeast, Central, and West—as shown in Table E-1. All 50 states and the District of Columbia are listed, with the participants in the Trial State Assessment shown in italic type. Territories were not assigned to a region. The part of Virginia that is included in the Washington, DC, metropolitan statistical area is included in the Northeast region; the remainder of the state is included in the Southeast region.

Table E-1
NAEP Geographic Regions

NORTHEAST	SOUTHEAST	CENTRAL	WEST
Connecticut	*Alabama*	Illinois	Alaska
Delaware	*Arkansas*	*Indiana*	*Arizona*
District of Columbia	*Florida*	*Iowa*	*California*
Maine	*Georgia*	Kansas	*Colorado*
Maryland	*Kentucky*	*Michigan*	*Hawaii*
Massachusetts	*Louisiana*	*Minnesota*	*Idaho*
New Hampshire	*Mississippi*	*Missouri*	Montana
New Jersey	*North Carolina*	*Nebraska*	Nevada
New York	*South Carolina*	*North Dakota*	*New Mexico*
Pennsylvania	*Tennessee*	*Ohio*	*Oklahoma*
Rhode Island	*Virginia*	South Dakota	Oregon
Vermont	*West Virginia*	*Wisconsin*	*Texas*
Virginia			*Utah*
			Washington
			Wyoming

MODAGE (Modal age)

The modal age (the age of most of the students in the grade sample) for the fourth grade students is age 9. A value of 1 for MODAGE indicates that the student is younger than the modal age; a value of 2 indicates that the student is of the modal age; a value of 3 indicates that the student is older than the modal age.

VARIABLES DERIVED FROM
THE STUDENT AND TEACHER QUESTIONNAIRES

Several variables were formed from the systematic combination of response values for one or more items from either the student demographic questionnaire, the student reading background questionnaire, or the teacher questionnaire.

HOMEEN2 (Home environment——Articles [of 4] in the home)

The variable HOMEEN2 was created from the responses to student demographic items B000901 (Does your family get a newspaper regularly?), B000903 (Is there an encyclopedia in your home?), B000904 (Are there more than 25 books in your home?), and B000905 (Does your family get any magazines regularly?). The values for this variable were derived as follows:

1 0-2 types	The student responded to at least two items and answered Yes to two or fewer.
2 3 types	The student answered Yes to three items.
3 4 types	The student answered Yes to four items.
8 Omitted	The student answered fewer than two items.

SINGLEP (How many parents live at home)

SINGLEP was created from items B005601 (Does either your mother or your stepmother live at home with you?) and B005701 (Does either your father or your stepfather live at home with you?). The values for SINGLEP were derived as follows:

1 2 parents at home	The student answered Yes to both items.
2 1 parent at home	The student answered Yes to B005601 and No to B005701, or Yes to B005701 and No to B005601.
3 Neither at home	The student answered No to both items.
8 Omitted	The student did not respond to or filled in more than one oval for one or both items.

TRCERTO (Teaching certificate - Other)

The variable TRCERTO was created from the responses to teacher background questions T040504 (Do you have teaching certification in middle/junior high school or secondary mathematics?) and T040505 (Do you have teaching certification in [some] other [category]?). The values for this variable were defined as follows:

1	Yes	Teacher indicated they were certified in mathematics or they were certified in "other."
2	No	Teacher indicated that they were not certified in either mathematics or "other."
3	Not offered	Teacher indicated that neither mathematics nor "other" certification was offered in their state.

TRUMAJB (Teacher undergraduate major - Other)

This variable was based on teachers' indications as to whether they had an undergraduate major in mathematics (T040703), mathematics education (T040704), or some other area (T040705).

1	Yes	Undergraduate major in mathematics, mathematics education, or "other."
2	No	No undergraduate major in mathematics, mathematics education, or "other."
8	Omitted	Teacher did not provide responses to any of the undergraduate majors.

TRGMAJB (Teacher graduate major - Other)

This variable was based on teachers' indications as to whether they had a graduate major in mathematics (T040803), mathematics education (T040804), or some other area (T040805).

1	Yes	Graduate major in mathematics, mathematics education, or "other."
2	No	No graduate major in mathematics, mathematics education, or "other."
8	Omitted	Teacher did not provide responses to any of the graduate majors.

TRTRAIN (Teacher training in reading activities)

This variable was created by examining teachers' responses to whether they had any training, in either college courses or in-service education, in eight areas:

T045903	Teaching critical thinking
T045904	The role of students' prior knowledge in their reading
T045907	Literature-based reading instruction
T045908	Reading assessment
T045909	Content area reading
T045910	Combining reading and writing
T045911	The whole language approach to teaching reading
T045912	Phonics in the teaching of reading

The number of times the teacher said "yes" to these questions was summed. The values of TRTRAIN were assigned as follows.

1	0-3	0 to 3 areas of training
2	4-6	4 to 6 areas of training
3	7	7 areas of training
4	8	8 areas of training

TREMP1 (Teaching heavy emphasis #1)

The responses to T046801 (How much instructional time is devoted to decoding skills?) and T046901 (How much emphasis is given to phonics?) were used to create TREMP1. If the teacher responded "almost all of the time" for T046801 and "heavy emphasis" for T046901, then TREMP1 was given a value of 1. Other combinations of responses were given a value of 2.

1	Yes	Heavy emphasis
2	No	No heavy emphasis
8	Omitted	Teacher omitted both questions

TREMP2 (Teacher heavy emphasis #2)

The responses to T046805 (How much instructional time is devoted to reading strategies?), T046902 (How much emphasis is given to the integration of reading and writing), and T046903 (How much emphasis is given to whole language?) were used to create TREMP2. The responses of "almost all of the time" for T046805 and "heavy emphasis" for T046902 and T046903 were considered. If the teacher provided the above responses to two or three of the questions, then TREMP2 was given a value of 1. Other combinations of responses were given a value of 2.

1	Yes	Heavy emphasis
2	No	No heavy emphasis
8	Omitted	Teacher omitted all three questions

TREMP3 (Teaching heavy emphasis #3)

The responses to T046904 (How much emphasis is given to reading across the content areas?) and T046905 (How much emphasis is given to individualized reading programs?) were used to create TREMP3. If the teacher responded "heavy emphasis" to T046904 and T046905, then TREMP3 was given a value of 1. Other combinations of responses were given a value of 2.

1	Yes	Heavy emphasis
2	No	No heavy emphasis
8	Omitted	Teacher omitted both questions

TRUMAJ (Teacher undergraduate major)

Items T040701 through T040705 in the teacher questionnaire (What were your undergraduate major fields of study?) were used to determine TRUMAJ as follows:

1	English/reading	The teacher responded yes to T040702 (English, reading, and/or language arts).
2	Education	The teacher responded yes to T040701 (education) and No to T040702.
3	Other	Any other response.

TRGMAJ (Teacher graduate major)

Items T040801 through T040806 in the teacher questionnaire (What were your undergraduate major fields of study?) were used to determine TRGMAJ as follows:

1	English/reading	The teacher responded yes to T040802 (English, reading, and/or language arts).
2	Education	The teacher responded yes to T040801 (education) and no to T040702.
3	Other	The teacher responded yes to T040803 (mathematics education), T040804 (mathematics), or T040805 (other).
4	None	The teacher indicated (T040806) that he or she had had no graduate-level study.

TRCERT (Type of teaching certification)

Items T040501 through T040505 (Do you have teaching certification in any of the following areas that is recognized by the state in which you teach?) were combined to create TRCERT. The following rules were used to determine the four values of TRCERT.

1	Reading	The teacher responded yes to T040502 (Reading)
2	Language arts	The teacher responded yes to T040503 (language arts) and no to T040502.
3	Education	The teacher responded yes to T040501 (education) and no to T040502 and T040503.
4	Other	Any other response

VARIABLES DERIVED FROM READING ITEMS

NORMIT (Normit Gaussian score)
SCHREAD (School-level mean Gaussian score)

The normit score is a student-level Gaussian score based on the inverse normal transformation of the mid-percentile rank of a student's number-correct booklet score within that booklet. The normit scores were used to decide collapsing of variables, finalize conditioning coding, and check the results of scaling.

The number correct is based on the number of dichotomous items answered correctly plus the score obtained on extended constructed-response items. The mid-percentile rank is based on the formula:

$$\frac{CF(i)+CF(i-1)}{2N}$$

where CF(i) is the cumulative frequency at i items correct and N is the total sample size. If i = 0 then

$$\frac{CF(0)+\frac{CF(1)}{2}}{2N}$$

A school-level normit, SCHREAD, was also created; this was the mean normit across all reading booklets administered in a school.

VARIABLES RELATED TO PROFICIENCY SCALING

Proficiency Score Variables

Item response theory (IRT) was used to estimate average reading proficiency for each state and for various subpopulations, based on students' performance on the set of reading items they received. IRT provides a common scale on which performance can be reported for the nation, state, and subpopulations, even when all students do not answer the same set of questions. This common scale makes it possible to report on relationships between students' characteristics (based on their responses to the background questions) and their overall performance in the assessment.

A scale ranging from 0 to 500 was created to report performance for each of the two content areas—Reading for Literary Experience and Reading to Gain Information. Each content-area scale was based on the distribution of student performance across all three grades assessed in the 1992 national assessment (grades 4, 8, and 12) and had a mean of 250 and a standard deviation of 50. A composite scale was created as an overall measure of students' mathematics proficiency. The composite scale for grade 4 was a weighted average of the two content area scales, where the weight for each content area was proportional to the relative importance assigned to the content area as specified in the mathematics objectives. Although the items comprising each scale were identical to those used for the national program, the item parameters for the Trial State Assessment scales were estimated from the combined data from all jurisdictions participating in the Trial State Assessment.

Scale proficiency estimates were obtained for all students assessed in the Trial State Assessment. The NAEP methods use random draws ("plausible values") from estimated proficiency distributions to compute population statistics. Plausible values are not optimal estimates of individual proficiency; instead, they serve as intermediate values to be used in estimating population characteristics. Chapter 8 provides further details on the computation and use of plausible values.

The proficiency score (plausible value) variables are provided on the student data files for each of the scales and are named as shown in Table E-2.

Table E-2
Scaling Variables for the 1992 Trial State Assessment Samples

Reading Scale	Data Variables
Reading for Literary Experience	RRPS11 to RRPS15
Reading to Gain Information	RRPS21 to RRPS25
Composite	RRPCM1 to RRPCM5

SMEANR (School mean score)
SNSCHR (Number of schools ranked)
SRANKR (School rank)
SRNK3R (Top, middle, bottom third)

A mean reading composite score (SMEANR) was calculated for each school included in the grade 4 assessment. The mean composite score was based on the values from the scaling variable RRPCM1 and was calculated using the students' sampling weights. The schools were then ordered from highest to lowest mean score within a jurisdiction—the school with the highest mean score was given a ranking of 1 and the school with the lowest mean score was given a ranking equal to the number of schools in the jurisdiction. Values for school rank are found in the variable SRANKR. The number of schools ranked (i.e., the number of schools in the jurisdiction with assessed students) is found in the variable SNSCHR.

These variables were later used in partitioning the schools within the national public-school comparison sample and the schools within each state into three groups based on their ranking (highest third, middle third, and lowest third). The data from the partitioning are found in the variable SRNK3R.

PRINCIPAL'S QUESTIONNAIRE VARIABLES (PQ)

Before the assessment, Westat, Inc., distributed a questionnaire to the principal of each participating school to gather data about school characteristics, including parents' occupations and student race/ethnicity. The data variables from this questionnaire are retained on the school file. A subset of these variables are also on the student files. Principal's questionnaire variables are identified in the data layouts by "(PQ)" in the SHORT LABEL field.

QUALITY EDUCATION DATA VARIABLES (QED)

The data files contain several variables obtained from information supplied by Quality Education Data, Inc. (QED). QED maintains and updates annually lists of schools showing grade span, total enrollment, instructional dollars per pupil, and other information for each school. These data variables are retained on both the school and student files and are identified in the data layouts by "(QED)" in the SHORT LABEL field.

Most of the QED variables are defined sufficiently in the data codebooks. Explanations of others are provided below.

ORSHPT and SORSHPT are the Orshansky Percentile, an indicator of relative wealth that specifies the percentage of school-age children in a district who fall below the poverty line.

IDP and SIDP represent, at the school district level, dollars per student spent for textbooks and supplemental materials.

ADULTED and SADLTED indicate whether or not adult education courses are offered at the school site.

URBAN and SURBAN define the school's urbanicity: urban (central city); suburban (area surrounding central city, but still located within the counties constituting the metropolitan statistical area); or rural (area outside any metropolitan statistical area).

APPENDIX F

THE NAEP ACHIEVEMENT LEVEL-SETTING PROCESS FOR THE 1992 READING ASSESSMENT

APPENDIX F

The NAEP Achievement Level-setting Process for the 1992 Reading Assessment

Mary Lyn Bourque[1]
National Assessment Governing Board

Introduction

Since 1984, NAEP has reported the performance of students in the nation and for specific subpopulations on a 0-to-500 proficiency scale. The history and development of the scale and the anchoring procedure used to interpret specific points on that scale are described in Appendix G.

Legislation[2] in 1988 created an independent board, the National Assessment Governing Board (NAGB), responsible for setting policy for the NAEP program. The Board has a statutory mandate to identify "appropriate achievement goals for each...grade in each subject area to be tested under the National Assessment." Consistent with this directive, and striving to achieve one of the primary mandates of the statute "to improve the form and use of NAEP results," the Board set performance standards (called achievement levels by NAGB) on the National Assessment in 1990, and again in 1992.

The 1990 trial, initiated in December 1989 with the dissemination of a draft policy statement (NAGB, 1989) and culminating 22 months later in the publication of the NAGB report, *The Levels of Mathematics Achievement* (Bourque & Garrison, 1991), consisted of two phases: the main study and a replication-validation study. Although there were slight differences between the two phases, there were many common elements. Both phases used a modified (iterative/empirical) Angoff (1971) procedure for arriving at the levels; both focused on estimating performance levels based on a review of the 1990 NAEP mathematics item pool; and both phases employed policy definitions for basic, proficient, and advanced levels (NAGB, 1990) as the criteria for rating items. The 1990 process was evaluated by a number of different groups (for a discussion, see Hambleton & Bourque, 1991) who identified technical flaws in the 1990 process. These evaluations influenced the Board's decision to set the levels again in 1992, and to not use the 1990 levels as benchmarks for progress toward the national goals during the coming decade. It is interesting to note, however, that the 1990 and 1992 processes produced remarkably similar results.

---

[1] The author is grateful to Susan Loomis, Richard Luecht, and Mel Webb, American College Testing, for their helpful suggestions and comments for improving earlier drafts of this paper.

[2] Public Law 100-297. (1988). National Assessment of Educational Progress improvement act (Article No. USC 1221). Washington, DC.

In September 1991, the Board contracted with American College Testing (ACT) to convene the panels of judges that would recommend the levels on the 1992 NAEP assessments in reading, writing, and mathematics. While the 1992 level-setting activities were not unlike those undertaken by the Board in 1990, there were significant improvements made in the process for 1992. There was a concerted effort to bring greater technical expertise to the process: the contractor selected by the Board has a national reputation for setting standards in a large number of certification and licensure exams; an internal and external advisory team monitored all the technical decisions made by the contractor throughout the process; and state assessment directors periodically provided their expertise and technical assistance at key stages in the project.

Setting achievement levels is a method for setting standards on the NAEP assessment that identify what students should know and be able to do at various points along the proficiency scale. The initial policy definitions of the achievement levels were presented to panelists along with an illustrative framework for more in-depth development and operationalization of the levels. Panelists were asked to determine descriptions/definitions of the three levels from the specific framework developed for the NAEP assessment with respect to the content and skills to be assessed. The operationalized definitions were refined throughout the level-setting process, as well as validated with a supplementary group of judges subsequent to the level-setting meetings. Panelists were also asked to develop a list of illustrative tasks associated with each of the levels, after which sample items from the NAEP item pool were identified to exemplify the full range of performance of the intervals between levels. The emphasis in operationalizing the definitions and in identifying and selecting exemplar items and papers was to represent the full range of performance from the lower level to the next higher level. The details of the implementation procedures are outlined in the remainder of this appendix.

Preparing for the Reading Level-setting Meeting

It is important for the planning of any standard-setting effort to know how various process elements interact with each other. For example, panelists interact with pre-meeting materials, meeting materials (i.e., the assessment questions, rating forms, rater feedback, and so forth), each other, and the project staff. All of these elements combine to promote or degrade what has been called intrajudge consistency and interjudge consensus (Friedman & Ho, 1990).

Previous research has conceptualized the effects of two major kinds of interaction: (1) people interacting with text (Smith & Smith, 1988), and (2) people interacting with each other Curry, 1987; Fitzpatrick, 1989). In order to assess the effects of textual and social interaction and adjust the standard-setting procedures accordingly, a pilot study was conducted as the first phase of the 1992 initiative.

Reading was chosen as the single content area to be pilot tested since it combined all of the various features found in the other NAEP assessments, including multiple-choice, and both short and extended constructed-response items. The pilot study provided the opportunity to implement and evaluate all aspects of the operational plan—background materials, meeting materials, study design, meeting logistics, staff function, and participant function.

2รบ

The overall pilot was quite successful. The level-setting process worked well, and the pilot allowed the contractor to make improvements in the design before implementation activities began. For example, schedule changes were made that allowed the panelists more time to operationalize the policy definitions before beginning the item-rating task. Also, the feedback mechanisms used to inform panelists about interjudge and intrajudge consistency data were improved for clarity and utility to the entire process.

The Reading Level-setting Panel

Sixty-four panelists representing 32 jurisdictions (31 states and the Virgin Islands) were selected from the 366 nominees and invited to participate in the level-setting process. They represented reading/language arts teachers at grades 4, 8, and 12, nonteacher educators, and members of the noneducator (general public) community. The group was balanced by gender, race/ethnicity, NAEP regions of the country, community type (low SES, not low SES), district size, and school type (public/private). Two panelists were unable to attend due to a family emergency and a loss of job, resulting in 62 participants, 22 at grade 4, 20 at grade 8, and 20 at grade 12.

Process for Developing the Achievement Levels

The four-and-one-half-day session began with a brief overview of NAEP and NAGB, a presentation on the policy definitions of the achievement levels, a review of the NAEP reading assessment framework, and a discussion of factors that influence item difficulty. The purpose of the presentation was to focus panelists' attention on the reading framework and to emphasize the fact that panelists' work was directly related to the NAEP assessment, not to the whole domain of reading.

All panelists completed and self-scored an appropriate grade-level form of the NAEP assessment. The purpose of this exercise was to familiarize panelists with the test content and scoring protocols—as well as time constraints—before beginning to develop the preliminary operationalized descriptions of the three levels.

Working in small groups of five or six, then eventually in grade-level groups, panelists expanded and operationalized the policy definitions of basic, proficient, and advanced in terms of specific reading skills, knowledge, and behaviors that were judged to be appropriate expectations for students in each grade, and to be in accordance with the current reading assessment framework.

The policy definitions are as follows:

Basic	This level, below proficient, denotes partial mastery of the knowledge and skills that are fundamental for proficient work at each grade—4, 8, and 12.
Proficient	This central level represents solid academic performance for each grade tested—4, 8, and 12. Students reaching this level have demonstrated competency

	over challenging subject matter and are well prepared for the next level of schooling.
Advanced	This higher level signifies superior performance beyond proficient grade-level mastery at grades 4, 8, and 12.

The small groups were allowed to brainstorm about what student performance *should be*, using the framework and their experience in completing the NAEP assessment as guides[3]. In addition, a practice task caused panelists to examine items in the half of the item pool that they would not be rating later. A comprehensive listing of grade level descriptors was developed, and panelists were asked to identify the five or six that best described what students *should be able to do* at each of the levels. Those descriptors appearing with the greatest frequency were compiled into a discussion list for the grade-level groups. Additions, deletions, and modifications were made as a result of discussions, and the groups reached general agreement that the final list of descriptors represented what students *should be able to do* at each achievement level.

Panelists next received two hours of training in the Angoff method. Training was customized to reflect the unique item formats of the particular subject area assessment. Once a conceptual consensus was reached about the characteristics of **marginally** acceptable examinees at each of the three levels, practice items from the released pool were rated by the panelists according to the process defined in the contractor's plan. For multiple-choice and short constructed-response items, panelists were asked to rate each item for the expected probability of a correct response for a group of *marginally* acceptable examinees at the basic, proficient, and advanced levels. For extended constructed-response items, panelists were asked to review a set of student response papers and select three papers, one for each achievement level, that typified *marginally* acceptable examinee performance for that level.

Following training in the Angoff method, the judges began the rating and paper selection process, inspecting and rating each dichotomously scored item in the pool for the expected probabilities of answering the item correctly at each level. For polytomously scored items, panelists reviewed a representative set of 24 to 28 student response papers for each item and selected the paper that best represented marginally acceptable student performance at each level. Panelists completed three rounds of item ratings and paper selections. For Round 1, panelists first answered the items related to a reading passage, then reviewed their answers using scoring keys and protocols. This process helped ensure that panelists would be thoroughly familiar with each item, including the foils and scoring rubrics, before rating the item. Panelists provided item ratings/paper selections for all three achievement levels, one item at a time, for all the items related to a reading passage, then proceeded to the next reading passage and set of items, for which the process was repeated. Panelists rated items for half the items in their grade-level assessment; one block of exercises was common to both halves of the grade-level groups. During Round 1, panelists used their lists of descriptors and other training materials for guidance in the rating process.

---

[3] The panelists also reviewed about half the item pool (the half they would not be rating later) so that the descriptors could be further modified if that was deemed appropriate.

Following Round 1, item response theory (IRT) was used to convert the rating results[4] for each rater to a latent ability scale, represented by the Greek letter theta (θ). This θ scale was the same scale to which the NAEP items evaluated by each panelist were calibrated. In order to provide meaningful feedback about item ratings, a special *relative scale* was constructed, which was a linear transformation of the theta scale having a mean of 75 and standard deviation of 15. Before Round 2 of the rating process, panelists were given interjudge consistency information using this relative scale. This information allowed panelists to see where their individual mean item ratings were on the scale, relative to the mean for the group and to the means for other panelists. Reasons for extreme mean ratings, including the possibility that some panelists misinterpreted the item rating task, were discussed briefly.

Before Round 2, panelists were also given item difficulty data. This information was presented as the percentage of students who answered each item correctly during the actual NAEP administration, for items scored "correct" or "incorrect" (i.e., multiple-choice and short constructed-response items), and as the mean score for student responses (on a scale of 1 to 4) for the extended response items. Panelists were told that this item difficulty information should be used as a reality check. For items on which item ratings differed substantially from the item difficulty value, panelists were asked to reexamine the item to determine if they had misinterpreted the item or misjudged its difficulty. Results of the data analysis, and panelists' own evaluations, indicated that the item difficulty information was perceived as very useful but had little impact on panelists' ratings.

For Round 2, panelists reviewed the same set of items they rated in Round 1 and, using the interjudge consistency information, the item difficulty information, and the information provided prior to Round 1, they either confirmed their initial item ratings and paper selections or adjusted their ratings to reflect the additional information. About one-half of Round 1 item ratings and paper selections were adjusted during Round 2.

Prior to Round 3, panelists' ratings were reanalyzed and additional information was presented to panelists concerning intrajudge variability. For each panelist, the intrajudge variability information consisted of those items that they had rated differently than items having similar difficulty, taking into consideration the panelist's aggregated item ratings. That is, the panelists' aggregated item ratings were converted to the theta (θ) scale. All items rated by the panelists were then analyzed in terms of the panelist's achievement level (θ) in comparison to actual student performance on the items. The observed item rating from each panelist was contrasted to an expected item rating. Those items with the largest differences between observed and expected ratings were identified. Panelists were given this information and asked to review each of these items and decide if their Round 2 ratings still accurately reflected their best judgments of the items. The intrajudge consistency data was to be used to flag items for reconsideration in the final round of rating.

For Round 3, panelists reviewed the same set of items they rated in Rounds 1 and 2 using both the new intrajudge variability information and the information made available during Rounds 1 and 2. In addition, panelists could discuss, within their small groups, ratings and paper selections for specific items about which they were unsure. About one-third of the item ratings were adjusted during Round 3.

---

[4] Because the IRT item parameters were not available for the polytomously scored (extended constructed-response) items, these items were not included in the following discussion of results.

Process of Selecting Exemplar Items

On the final day of the achievement level-setting process, panelists reviewed items from the 1992 item pool scheduled for release to the public. The released item pool was the set from which the panelists could select items illustrative of the achievement levels for their grade. Exercises are organized in blocks, consisting of a reading passage, followed by several items, usually employing each of the three item formats, (i.e., multiple-choice, short constructed-response, and extended constructed-response). A total of 10 blocks from the 1992 exercise pool were scheduled for release: 2 blocks from the fourth-grade pool, totaling 19 items; 4 blocks from the eighth-grade pool, totaling 52 items; and 4 blocks from the twelfth-grade pool, totaling 46 items.

Panelists who had rated specific blocks of released items were asked to review those same items again to select particular ones as exemplary of each achievement level. The items were pre-assigned to each achievement level based on the final round of the judges' rating data, and using the following statistical criteria. For any given level (basic, proficient, or advanced),

(1) items having an expected p-value[5] $\geq .501$ and $\leq .750$, *at that level,* were assigned to that level;

(2) items meeting the criteria at *more than one level* were assigned to *one* level taking both the expected p-value and the appropriateness of the item for one of the levels into account; and

(3) items with expected p-values $\leq .501$ were assigned to levels where a specific passage had few or no items at that level.

For example, the raters' expected p-value for one of the released items might have been .366 at the basic level, .701 at the proficient level, and .932 at the advanced level. This item would have been identified for review as a potential exemplar item for the proficient level. The expected p-value at the basic level was too low for consideration as a basic-level exemplar—that is, the item was judged to be too difficult, and the expected p-value at the advanced level was too high for consideration at the advanced level—that is, the item was judged to be too easy. Table F-1 shows the results of this process for each grade and level.

Panelists were asked to review the items as classified, and form an individual judgment regarding the suitability of each item to illustrate and further communicate the meaning of the levels. Each item's classification could be accepted, rejected, or reassigned, although the

---

[5] Expected p-values were based on the average predicted performance at the cut point for each achievement level.

Table F-1

Results of First Review for Achievement-level Exemplars

Level/Status	Grade 4	Grade 8	Grade 12	All Grades
Total released	19	52	46	117
Basic				
Reviewed	4	12	18	34
Recommended	3	5	14	22
Proficient				
Reviewed	5	14	20	39
Recommended	4	12	9	25
Advanced				
Reviewed	5	6	7	18
Recommended	5	6	8	19

Table F-2

Results of Review of Additional Items for Achievement-level Exemplars

Level/Status	Grade 4	Grade 8	Grade 12	All Grades
Total items recommended	13	13	21	47
Basic				
Reviewed	3	12	12	27
Recommended	6	7	8	21
Proficient				
Reviewed	4	13	11	28
Recommended	6	3	8	17
Advanced				
Reviewed	5	8	9	22
Recommended	1	3	5	9

procedure was primarily designed to eliminate items that did not meet panelists' expectations for any reason. Items were reclassified if a strong consensus was found to hold for that change.

During the validation process, described in the next section, items were again reviewed. Those that had been selected by the original standard-setting panel were grouped into sets of *pre-selected* items. All remaining items in the released blocks that met the statistical criteria, *but were not recommended by the original panel*, were grouped into a set identified as *additional items for review*. Exercises that had been recommended for reclassification into another achievement level category were presented in their original classification for purposes of this review. As the Table F-2 shows, 21 items were recommended as exemplars for the basic level, 17 for the proficient level, and 9 for the advanced.

Process for Validating the Levels

Nineteen reading educators participated in the item selection and content validation process. Ten of the panelists were reading teachers who had participated in the original achievement level-setting process and who had been identified as outstanding panelists by grade group facilitators during this meeting, who were extensively involved with professional organizations (e.g., the International Reading Association, the National Reading Conference, or the National Council for Teachers of English), and who had outstanding service credentials. The other nine panelists represented state-level reading curriculum supervisors or assessment directors, as well as university faculty teaching in disciplines related to this subject area. To the extent possible, the group was balanced by race/ethnicity and gender.

The two-and-one-half day meeting began by briefing panelists on the purpose of the meeting and by giving them an overview of the level-setting process and results. Panelists first reviewed the operationalized descriptions of the achievement levels for qualities such as (1) within- and across-grade consistency, (2) grade-level appropriateness, and (3) utility for increasing the public's understanding of the NAEP reading results. Next, panelists reviewed the operationalized descriptions of the achievement levels for consistency with the NAGB policy definitions of basic, proficient, and advanced with the NAEP *Reading Objectives*. Working in grade-level (4, 8, and 12) groups of 6 to 7 panelists each, then as a whole group, panelists reviewed the operationalized descriptions to provide within- and across-grade consistency, and to align the language and concepts of the descriptions more closely with the language of the NAEP *Reading Objectives*. (Both the original descriptions and the revised descriptions are included later in this appendix.) Finally, panelists suggested revisions they thought would improve the operational descriptions based on their earlier reviews.

On the final day, panelists worked in grade-level groups to review the possible exemplar items. The task was to select a set of items, for each achievement level for their grade, that would best communicate to the public the levels of reading ability and the types of skills needed to perform in reading at that level.

After selecting sets of items for their grades, the three grade-level groups met as a whole group to review item selection. During this process, cross-grade items that had been selected as exemplars for two grades (two such items were selected for grades 8 and 12) were assigned to one grade by whole-group consensus. In addition, items were evaluated by the whole group for overall quality. This process yielded 13 items as recommended exemplars for grade 4, 13 items as recommended exemplars for grade 8, and 21 items as recommended exemplars for grade 12.

Mapping the Levels onto the NAEP Scale

The process of mapping panelists' ratings to the NAEP scales used *item response theory* (IRT). IRT provided statistically sophisticated methods for determining the expected performance of examinees on particular test items in terms of an appropriate measurement scale. The same measurement scale simultaneously described the characteristics of the test items and the performance of the examinees. Once the item characteristics were set, it was possible to determine precisely how examinees were likely to perform on the test items at different points of the measurement scale.

The panelists' ratings of the NAEP test items were likewise linked, by definition, to the expected performance of examinees at the theoretical achievement level cut points. It was therefore feasible to use the IRT item characteristics to calculate the values on the measurement scale corresponding to each achievement level. This was done by averaging the item ratings over panelists for each achievement level and then simply using the item characteristics to find the corresponding achievement level cut points on the IRT measurement scale. This process was repeated for each of the NAEP reading scales within each grade (4, 8, and 12).

For the multiple-choice and short constructed-response items that were dichotomously scored, the judges each rated half of the items in the NAEP pool in terms of the expected probability that a student at a borderline achievement level would answer the item correctly, based on the judges' operationalization of the policy definitions and the factors that influence item difficulty. To assist the judges in generating consistently scaled ratings, the rating process was repeated twice, with feedback. Information on consistency among different judges and on the difficulty of each item[6] was fed back into the first repetition (Round 2), while information on consistency within each judge's set of ratings was fed back into the second repetition (Round 3). The third round of ratings permitted the judges to discuss their ratings among themselves to resolve problematic ratings. The mean final rating of the judges aggregated across multiple-choice and short constructed-response items yielded the threshold values for these items in the percent correct metric. These cut scores were then mapped onto the NAEP scale (which is defined and scored using item response theory, rather than percent correct).

For extended constructed-response items, judges were asked to select student papers that exemplified performance at the cutpoint of each achievement level. Then for each achievement level, the mean of the scores assigned to the selected papers was mapped onto the NAEP scale in a manner similar to that used for the items scored dichotomously.

The final cut score for each achievement level was a weighted average of the cut score for the multiple-choice and short constructed-response items and the cut score for the extended constructed-response items, with the weights being proportional to the information supplied by the two classes of items. The judges' ratings, in both metrics, and their associated errors of measurement are shown in Table F-3.

---

[6] Item difficulty estimates were based on a preliminary, partial set of responses to the national assessment.

Table F-3
Cutpoints for Achievement Levels

Level	Mean Percent Correct, Multiple-choice and Short Constructed-response (Round 3)	Mean Paper Rating, Extended Constructed-response (Round 3)	Scale Score*	Standard Error of Scale Score**
Grade 4				
Basic	38	2.72	212	2.5
Proficient	62	3.14	243	2.1
Advanced	80	3.48	275	8.8
Grade 8				
Basic	41	2.13	244	2.6
Proficient	66	2.66	283	0.8
Advanced	85	3.22	328	7.7
Grade 12				
Basic	41	2.42	269	7.9
Proficient	67	2.85	304	2.8
Advanced	86	3.14	348	4.1

---

*Scale score is derived from a weighted average of the mean percents correct for multiple-choice and short constructed-response items and the mean paper ratings for extended constructed-response items after both were mapped onto the NAEP scale.

**The standard error of the scale is estimated from the difference in mean scale scores for the two equivalent subgroups of judges.

In the final stage of the mapping process, the achievement level cut points on the IRT measurement scale were combined over content areas and rescaled to the NAEP score scale. Weighted averages of the achievement level cut points were computed. The weighting constants accounted for the measurement precision of the test items evaluated by the panelists, the proportion of items belonging to each NAEP content area, and the linear NAEP scale transformations. These weighted averages produced the final cut points for the basic, proficient, and advanced achievement levels within each grade.

Figure F-1

Final Descriptions of 1992 Reading Achievement Levels

PREAMBLE

Reading for meaning involves a dynamic, complex interaction between and among the reader, the text, and the context. Readers, for example, bring to the process their prior knowledge about the topic, their reasons for reading it, their individual reading skills and strategies, and their understanding of differences in text structures.

The texts used in the reading assessment are representative of common real world reading demands. Students at grade 4 are asked to respond to literary and informational texts which differ in structure, organization, and features. Literary texts include short stories, poems, and plays that engage the reader in a variety of ways, not the least of which is reading for fun. Informational texts include selections from textbooks, magazines, encyclopedias, and other written sources whose purpose is to increase the reader's knowledge.

In addition to literary and informational texts, students at grades 8 and 12 are asked to respond to practical texts (e.g., bus schedules or directions for building a model airplane) that describe how to perform a task.

The context of the reading situation includes the purposes for reading that the reader might use in building a meaning of the text. For example, in reading for literary experience, students may want to see how the author explores or uncovers experiences, or they may be looking for vicarious experience through the story's characters. On the other hand, the student's purpose in reading informational texts may be to learn about a topic (such as the Civil War or the oceans) or to accomplish a task (such as getting somewhere, completing a form, or building something).

The assessment asks students at all three grades to build, extend, and examine text meaning from four stances or orientations:

> **Initial Understanding**—Students are asked to provide the overall or general meaning of the selection. This includes summaries, main points, or themes.
>
> **Developing Interpretation**—Students are asked to extend the ideas in the text by making inferences and connections. This includes making connections between cause and effect, analyzing the motives of characters, and drawing conclusions.
>
> **Personal Response**—Students are asked to make explicit connections between the ideas in the text and their own background knowledge and experiences. This includes comparing story characters with themselves or people they know, for example, or indicating whether they found a passage useful or interesting.
>
> **Critical Stance**—Students are asked to consider how the author crafted a text. This includes identifying stylistic devices such as mood and tone.

Figure F-1 (continued)

Final Descriptions of 1992 Reading Achievement Levels

These stances are not considered hierarchical or completely independent of each other. Rather, they provide a frame for generating questions and considering student performance at all levels. All students at all levels should be able to respond to reading selections from all of these orientations. What varies with students' developmental and achievement levels is the amount of prompting or support needed for response, the complexity of the texts to which they can respond, and the sophistication of their answers.

INTRODUCTION

The following achievement-level descriptions focus on the interaction of the reader, the text, and the context. They provide some specific examples of reading behaviors that should be familiar to most readers of this document. The specific examples are not inclusive; their purpose is to help clarify and differentiate what readers performing at each achievement level should be able to do. While a number of other reading achievement indicators exist at every level, space and efficiency preclude an exhaustive listing.

It should also be noted that the achievement levels are cumulative from basic to proficient to advanced. One level builds on the previous levels such that knowledge at the proficient level presumes mastery of the basic level, and knowledge at the advanced level presumes mastery at both the basic and proficient.

Grade 4—Basic

Fourth-grade students performing at the **basic level** *should demonstrate an understanding of the overall meaning of what they read.* When reading texts appropriate for fourth graders, *they should be able to make relatively obvious connections between the text and their own experiences.*

For example, when reading **literary text**, they should be able to tell what the story is generally about—providing details to support their understanding—and be able to connect aspects of the stories to their own experiences.

When reading **informational text**, basic-level fourth graders should be able to tell what the selection is generally about or identify the purpose for reading it; provide details to support their understanding; and connect ideas from the text to their background knowledge and experiences.

Grade 4—Proficient

Fourth grade students performing at the **proficient level** *should be able to demonstrate an overall understanding of the text, providing inferential as well as literal information.* When reading text

Figure F-1 (continued)

Final Descriptions of 1992 Reading Achievement Levels

appropriate to fourth grade, *they should be able to extend the ideas in the text by making inferences, drawing conclusions, and making connections to their own experiences. The connection between the text and what the student infers should be clear.*

For example, when reading **literary text**, proficient-level fourth graders should be able to summarize the story, draw conclusions about the characters or plot, and recognize relationships such as cause and effect.

When reading **informational text**, proficient-level students should be able to summarize the information and identify the author's intent or purpose. They should be able to draw reasonable conclusions from the text, recognize relationships such as cause and effect or similarities and differences, and identify the meaning of the selection's key concepts.

Grade 4—Advanced

Fourth-grade students performing at the **advanced level** *should be able to generalize about topics in the reading selection and demonstrate an awareness of how authors compose and use literary devices.* When reading text appropriate to fourth grade, *they should be able to judge texts critically and, in general, give thorough answers that indicate careful thought.*

For example, when reading **literary text**, advanced-level students should be able to make generalizations about the point of the story and extend its meaning by integrating personal experiences and other readings with the ideas suggested by the text. They should be able to identify literary devices such as figurative language.

When reading **informational text**, advanced-level fourth graders should be able to explain the author's intent by using supporting material from the text. They should be able to make critical judgments of the form and content of the text and explain their judgments clearly.

Grade 8—Basic

Eighth-grade students performing at the **basic level** *should demonstrate a literal understanding of what they read and be able to make some interpretations.* When reading text appropriate to eighth grade, *they should be able to identify specific aspects of the text that reflect the overall meaning, recognize and relate interpretations and connections among ideas in the text to personal experience, and draw conclusions based on the text.*

For example, when reading **literary text**, basic-level eighth graders should be able to identify themes and make inferences and logical predictions about aspects such as plot and characters.

When reading **informative text**, they should be able to identify the main idea and the author's purpose. They should make inferences and draw conclusions supported by information in the text.

Figure F-1 (continued)

Final Descriptions of 1992 Reading Achievement Levels

They should recognize the relationships among the facts, ideas, events, and concepts of the text (e.g., cause and effect and chronological order).

When reading **practical text**, they should be able to identify the main purpose and make predictions about the relatively obvious outcomes of procedures in the text.

Grade 8—Proficient

Eighth-grade students performing at the **proficient level** *should be able to show an overall understanding of the text, including inferential as well as literal information.* When reading text appropriate to eighth grade, *they should extend the ideas in the text by making clear inferences from it, by drawing conclusions, and by making connections to their own experiences—including other reading experiences.* Proficient eighth graders *should be able to identify some of the devices authors use in composing text.*

For example, when reading **literary text**, students at the proficient level should be able to give details and examples to support themes that they identify. They should be able to use implied as well as explicit information in articulating themes; to interpret the actions, behaviors, and motives of characters; and to identify the use of literary devices such as personification and foreshadowing.

When reading **informative text**, they should be able to summarize the text using explicit and implied information and support conclusions with inferences based on the text.

When reading **practical text**, proficient-level students should be able to describe its purpose and support their views with examples and details. They should be able to judge the importance of certain steps and procedures.

Grade 8—Advanced

Eighth-grade students performing at the **advanced level** *should be able to describe the more abstract themes and ideas of the overall text.* When reading text appropriate to eighth grade, they *should be able to analyze both meaning and form a[illegible]port their analyses explicitly with examples from the text; they should be able to extend text information by relating it to their experiences and to world events.* At this level, student *responses should be thorough, thoughtful, and extensive.*

For example, when reading **literary text**, advanced-level eighth graders should be able to make complex, abstract summaries and theme statements. They should be able to describe the interactions of various literary elements (i.e., setting, plot, characters, and theme); to explain how the use of literary devices affects both the meaning of the text and their response to the author's style. They should be able critically to analyze and evaluate the composition of the text.

Figure F-1 (continued)

Final Descriptions of 1992 Reading Achievement Levels

When reading **informative text**, they should be able to analyze the author's purpose and point of view. They should be able to use cultural and historical background information to develop perspectives on the text and be able to apply text information to broad issues and world situations.

When reading **practical text**, advanced-level students should be able to synthesize information that will guide their performance, apply text information to new situations, and critique the usefulness of the form and content.

Grade 12—Basic

Twelfth-grade students performing at the **basic level** *should be able to demonstrate an overall understanding and make some interpretations of the text.* When reading text appropriate to twelfth grade, they *should be able to identify and relate aspects of the text to its overall meaning, recognize interpretations, make connections among and relate ideas in the text to their personal experiences, and draw conclusions.* They *should be able to identify elements of an author's style.*

For example, when reading **literary text**, twelfth-grade students should be able to explain the theme, support their conclusions with information from the text, and make connections between aspects of the text and their own experiences.

When reading **informational text**, basic-level twelfth graders should be able to explain the main idea or purpose of a selection and use text information to support a conclusion or make a point. They should be able to make logical connections between the ideas in the text and their own background knowledge.

When reading **practical text**, they should be able to explain its purpose and the significance of specific details or steps.

Grade 12—Proficient

Twelfth-grade students performing at the **proficient level** *should be able to show an overall understanding of the text, which includes inferential as well as literal information.* When reading text appropriate to twelfth grade, they *should be able to extend the ideas of the text by making inferences, drawing conclusions, and making connections to their own personal experiences and other readings. Connections between inferences and the text should be clear, even when implicit.* These *students should be able to analyze the author's use of literary devices.*

When reading **literary text**, proficient-level twelfth graders should be able to integrate their personal experiences with ideas in the text to draw and support conclusions. They should be able to explain the author's use of literary devices such as irony or symbolism.

Figure F-1 (continued)

Final Descriptions of 1992 Reading Achievement Levels

When reading **informative text**, they should be able to apply text information appropriately to specific situations and integrate their background information with ideas in the text to draw and support conclusions.

When reading **practical texts**, they should be able to apply information or directions appropriately. They should be able to use personal experiences to evaluate the usefulness of text information.

Grade 12—Advanced

Twelfth-grade students performing at the **advanced level** *should be able to describe more abstract themes and ideas in the overall text.* When reading text appropriate to twelfth grade, they *should be able to analyze both the meaning and the form of the text and explicitly support their analyses with specific examples from the text.* They *should be able to extend the information from the text by relating it to their experiences and to the world.* Their *responses should be thorough, thoughtful, and extensive.*

For example, when reading **literary text**, advanced-level twelfth graders should be able to produce complex, abstract summaries and theme statements. They should be able to use cultural, historical, and personal information to develop and explain text perspectives and conclusions. They should be able to evaluate the text, applying knowledge gained from other texts.

When reading **informational text**, they should be able to analyze, synthesize, and evaluate points of view. They should be able to identify the relationship between the author's stance and elements of the text. They should be able to apply text information to new situations and to the process of forming new responses to problems or issues.

When reading **practical texts,** advanced-level twelfth graders should be able to make a critical evaluation of the usefulness of the text and apply directions from the text to new situations.

Figure F-2

Draft Descriptions of the Achievement Levels Prepared by the Original Level-setting Panel

4th-Grade Draft Descriptions

BASIC performance in reading should include:

- Determining what a text is about
- Identifying characterizations, settings, conflicts, or plots in a story
- Supporting one's understanding of a text with appropriate details
- Explaining why one likes or dislikes a text
- Connecting material in a text to personal experiences
- Making predictions about situations beyond the confines of a text
- Demonstrating an ability to maintain a focus over the entirety of a longer text

PROFICIENT performance in reading should include:

- Summarizing a text
- Recognizing an author's intent or purpose
- Making simple inferences based on information provided in a text
- Using information from a text to draw a basic conclusion
- Determining the meaning of key concepts in the text and connecting them to the main idea
- Recognizing the progression of ideas and the cause-and-effect relationships in a text
- Using the surrounding text to assign meaning to a word or phrase

ADVANCED performance in reading should include:

- Explaining an author's intent, using supporting material from the text
- Describing the similarities and differences in characters
- Demonstrating an awareness of the use of literary devices and figurative language
- Applying inferences drawn from a text to personal experiences
- Extending the meaning of a text by integrating experiences and information outside of the text
- Making and explaining a critical judgment of a text
- Demonstrating an ability to adapt reading purpose to genre and/or writing style

8th-Grade Draft Descriptions

BASIC performance in reading should include:

- Identifying the main idea or purpose of a text using information both stated and implied
- Expressing an author's purpose, viewpoint, and/or theme

Figure F-2 (continued)

Draft Descriptions of the Achievement Levels Prepared by the Original Level-setting Panel

- Using information from a text to draw and support conclusions
- Making inferences appropriate to the information provided in a text
- Recognizing the cause-and-effect relationships in a text
- Making logical connections from the material in a text to personal knowledge and experience

PROFICIENT performance in reading should include:

- Restating the main idea using supportive details and examples from a text
- Summarizing a text using information both stated and implied
- Making inferences from a text in order to draw valid conclusions
- Interpreting the actions, behaviors, and motives of characters
- Integrating personal knowledge and experience to enhance one's understanding of a text
- Identifying an author's use of literary devices

ADVANCED performance in reading should include:

- Describing how specific literary elements interact with each other
- Synthesizing the information in a text to obtain abstract meaning or to perform a task
- Finding new applications for information derived from a text
- Making personal and critical evaluations of a text
- Analyzing an author's purpose, viewpoint, and/or theme
- Explaining an author's use of literary devices

12th-Grade Draft Descriptions

BASIC performance in reading should include:

- Explaining the main idea of a text
- Describing the main purpose in reading a selection
- Recognizing the significance of details from a reading in order to support a conclusion or perform a task
- Applying the information gathered from reading to meet an objective or support a conclusion
- Explaining the basic elements of an author's literary devices

PROFICIENT performance in reading should include:

- Drawing conclusions from and making inferences about information from different texts and writing styles
- Integrating background information with newly acquired information to support conclusions

Figure F-2 (continued)

Draft Descriptions of the Achievement Levels
Prepared by the Original Level-setting Panel

- Applying information from a text in an appropriate manner
- Bringing personal experience and accumulated knowledge into the process of critically evaluating a text
- Explaining an author's purpose in using complex literary devices

ADVANCED performance in reading should include:

- Providing innovative elaborations from textual information
- Analyzing and evaluating different points of view by means of comparison and contrast
- Identifying the relationships between an author's or narrator's stance and the various elements of the text
- Critically evaluating a text within a specific frame of reference
- Bringing the knowledge of other texts to the process of critical evaluation
- Using cultural or historical information provided in a text to develop perspectives on other situations
- Using cultural or historical information to develop perspectives on a text

Figure F-3

Revised Draft Descriptions of the Achievement Levels Recommended by the Follow-up Validation Panel

Revised 4th-Grade Draft Descriptions

Basic performance in reading should include:

- Determining what a story/informational text is about (i.e. topic, main idea)
- Determining the main purpose for reading a selection
- Identifying character(s), setting(s), conflict(s), or plot(s) in a story
- Supporting one's understanding of a story/informational text with appropriate details
- Explaining why one likes or dislikes what they have read [a reading]
- Connecting material from a story/informational text to personal experiences
- Making predictions about situations beyond the confines of the printed material
- Maintaining a focus over the entirety of a story/informational text

Proficient performance in reading should include:

- Summarizing a story/informational text
- Recognizing an author's intent or purpose
- Making simple inferences based on information provided in a story/informational text
- Drawing a valid conclusion from a story/informational text
- Determining the meaning of key concepts in the story/informational text and connecting them to the main idea
- Recognizing relationships in a story/informational text (time order, cause/effect, compare/contrast)

Advanced performance in reading should include:

- Explaining an author's intent, using supporting material from the story/informational text
- Describing the similarities and difference in characters, settings, and plots
- Demonstrating an awareness of the use of literary devices, such as figurative language
- Applying inferences drawn from a story/informational text to personal experiences
- Extending the meaning of a story/informational text by integrating experiences and information outside of the text
- Making and explaining a critical judgment of a story/informational text
- Demonstrating an ability to adapt reading purpose to a variety of printed material and/or writing style

Revised 8th-Grade Draft Descriptions

Basic performance in reading should include:

- Identifying the main idea, theme, or purpose of a text
- Describing the main purpose for reading a selection

Figure F-3 (continued)

Revised Draft Descriptions of the Achievement Levels Recommended by the Follow-up Validation Panel

- Expressing an author's purpose and viewpoint
- Making inferences, predictions, and drawing conclusions that are supported by information in a text
- Recognizing the relationships among facts, ideas, events, and concepts within a text (e.g., cause and effect, chronological order, and characterization)
- Making logical connections between the text and personal knowledge
- Maintaining a focus over the entirety of a story/informational text

Proficient performance in reading should include:

- Restating the main idea, theme, or purpose of a text using supporting details and examples
- Summarizing a text using both stated and implied information
- Interpreting the actions, behaviors, and motives of characters
- Using personal knowledge and experience to enhance one's understanding of a text
- Identifying an author's use of literary devices (i.e. personification, foreshadowing, and so forth).
- Using inferences from a text in order to draw valid conclusions.

Advanced performance in reading should include:

- Describing how specific literary elements (i.e., setting, plot, characters, and theme) interact with each other
- Synthesizing the information in a text to obtain implied meaning or to perform a task
- Applying information derived from a text to new situations.
- Explaining an author's use of literary devices (i.e., irony, personification, and foreshadowing)
- Responding personally and critically to a text
- Analyzing an author's purpose and viewpoint
- Using cultural or historical information to develop perspectives on a text
- Using cultural or historical information provided in a text to develop perspectives on other situations

Revised 12th-Grade Draft Descriptions

Basic performance in reading should include:

- Explaining the main idea, theme, or purpose of a text
- Describing the main purpose for reading a selection
- Recognizing the significance of details from a reading in order to support a conclusion or perform a task

Figure F-3 (continued)

Revised Draft Descriptions of the Achievement Levels
Recommended by the Follow-up Validation Panel

- Applying the information gathered from reading to meet an objective or support a conclusion
- Identifying and explaining the basic elements of an author's literary devices
- Making logical connections between a text and personal knowledge and experience
- Maintaining a focus over the entirety of a story/informational text

Proficient performance in reading should include::

- Drawing conclusions and making inferences from different texts and writing styles
- Integrating background information with newly acquired information to support conclusions
- Applying information from a text in an appropriate manner
- Applying personal experience and accumulated knowledge to the process of critically evaluating a text
- Explaining an author's purpose in using complex literary devices (i.e. irony, symbolism)

Advanced performance in reading should include:

- All basic and proficient reading behaviors listed previously
- Prompted by information from a text, innovating in new situations and creating new answers to old situations
- Analyzing, synthesizing, and evaluating different points of view by means of comparison and contrast
- Identifying the relationships between an author's or narrator's stance and the various elements of the text
- Critically evaluating a text within a frame of reference
- Applying the knowledge of other texts to the process of critical evaluation
- Using cultural or historical information to develop perspectives on a text
- Using cultural or historical information provided in a text to develop perspectives on other situations

Figure F-4

Meeting Participants, NAEP Reading Achievement Level Setting
Original Meeting, St. Louis, Missouri, August 21 - 25, 1992

David Awbrey
Wichita Eagle
Wichita, KS

Dorothy Botham
Milwaukee Public Library
Milwaukee, WI

Anna Caballero
Attorney
Salinas, CA

Kathy Casseday
WFSP Radio Station
Kingwood, WV

Dee Ellis
Trimble Banner Newspaper
Milton, KY

Nona Smith
NAACP
New York, NY

Lillaine Speese
Oakdale Elementary School
Oroville, CA

Clifton Whetten
Retired Construction Sprvsr.
Elfrida, AZ

P. Richard Brackett
Brackett & Assoc. Motivational Marketing
Company
Brentwood, TN

Kathleen Harkey
Corporate Presentations
Nashville, TN

Patricia Oliverez
Salinas Public Library
Salinas, CA

Christine Sentz
North Milwaukee Branch Library
Milwaukee, WI

Carolyn Sullivan
Planters & Merchants Bank
Gillett, AR

Paula Abrams
City Hall
Bedford, KY

Rhonda Cantrell Dunn
Nashville Urban League
Nashville, TN

Harlon Gaskill (CPA)
Gaskill, Pharis & Pharis
Dalhart, TX

Jean McManis
Local/State Education Volunteer
State College, PA

Linda Borsum
Lakeview School District
Battlecreek, MI

Anne Kraut
Elementary Supervisor
Princeton, WV

Robert Williams
Macomb Intermediate SD
Clinton Township MI

Figure F-4 (continued)

Meeting Participants, NAEP Reading Achievement Level Setting
Original Meeting, St. Louis, Missouri, August 21 - 25, 1992

Constance Boyd
Owen J. Roberts SD
King of Prussia, PA

Mary Gonzalez
Mesa Public Schools
Mesa, AZ

James Schindler
Jordan SD
Salt Lake City, UT

Kathryn Flannery
Indiana University
Bloomington, IN

Catherine Hatala
School District of Philadelphia
Philadelphia, PA

Raymond Morgan
Old Dominion University
Virginia Beach, VA

Berton Wiser
Columbus Public School
Columbus, OH

Freda Andrews
Durham Public Schools
Durham, NC

Tim Barnes
Ashdown Public Schools
Ashdown, AR

Larry Barretto
Maplewood Elementary School
Coral Springs, FL

Gloria Darling
Conway Public Schools
Conway, AR

Nina Frederick
Marion County School System
Hackleburg, AL

Karen Fugita
Oak Grove SD
San Jose, CA

Anne Gregory
Durham Public Schools
Durham, NC

Joseph Howard
Josiah Quincy School
West Roxbury, MA

Roberta Johnson
Cleveland Public Schools
Cleveland, OH

Marcia Jolicoeur
Lisbon Falls School
Lewiston, ME

Elizabeth Litchfield
Westwood School District
Emerson, NJ

Jean Young
Houston ISD
Houston, TX

Wilma Centers
Wolfe County Middle School
Campton, KY

Eunice Coakley
Greenville School
Greenville, SC

Eugenia Constantinou
Prince Georges County Schools
Silver Spring, MD

Figure F-4 (continued)

Meeting Participants, NAEP Reading Achievement Level Setting
Original Meeting, St. Louis, Missouri, August 21 - 25, 1992

Walt Cottingham
Henderson City Schools
Zirconia, NC

Deborah Davidson
Westhampton Beach UFSD
Patchogue, NY

Julia Dominique
Department of Education USVI
Sunnyisle, VI

Patricia Gerdes
Waelder ISD
Schulenburg, TX

Leslie Leech
Elkton School
Elkton, SD

Belva Leffel
Whittier Christian Jr. High
Norwalk, CA

Harriett McAllaster
Volusia County Schools
DeLand, FL

Mary Orear
Camden-Rockport HS & MS
Rockport, ME

Judith Zinsser
Houston ISD
Houston, TX

Mary Ann Ledbetter
East Baton Rouge Parish School Board
Baton Rouge, LA

Cora Cummins
Conway Public Schools
Conway, AR

Stanley Fraundorf
Cuba City Public Schools
Cuba City, WI

Georgia Howard
Volusia County Schools
Holly Hill, FL

Roger Larsen
Campbell County SD
Gillette, WY

Judith Lusk
Norfield School District
Rockbury, VT

Donnie McQuinn
Wolfe County Board of Education
Pine Ridge, KY

Meredith Powers
Swansea School
Providence, RI

Beth Schieber
Kingfisher Schools
Okarche, OK

Carolyn Sue Wilson
Greenville, SC

Sue Zak
Cleveland Board of Education
Garfield Heights, OH

Figure F-5

Meeting Participants, NAEP Reading Achievement Level Setting Follow-Up Validation Meeting, San Diego, California, October 9 - 11, 1992

Meredith Powers
Swansea School
Providence, RI

Roger Larsen
Campbell County SD
Gillett, WY

Beth Schieber
Kingfisher Schools
Okarche, OK

Elizabeth Litchfield
Westwood School District
Emmerson, NJ

Larry Barretto
Maplewood Elementary School
Coral Springs, FL

Anne Gregory
Durham Public Schools
Durham, NC

Debra Davidson
Westhampton Beach UFSD
Patchogue, NY

Eugenia Constantinou
Prince Georges County School
Silver Spring, MD

Eunice Coakley
Greenville School
Greenville, SC

Nancy Livingston
Brigham Young University
Salt Lake City, UT

Susan McIntyre
University Wisconsin-Eau Claire
Eau Claire, WI

Clyde Colwell
Norfolk Public School
Norfolk, VA

Jo Prather
Mississippi Department of Education
Jackson, MS

Mary Orear
Camden-Rockport HS & MS
Rockport, ME

Shelia Potter
Michigan Department of Education
Lansing, MI

Gene Jongsma
IRA Subcommittee Member
San Antonio, TX

Peggy Dutcher
Michigan Education Assessment Program
Lansing, MI

Martha Carter
Milwaukee Public Schools
Milwaukee, WI

Mark Conley
Michigan State University
Holt, MI

APPENDIX G

THE NAEP SCALE ANCHORING PROCESS FOR THE 1992 READING ASSESSMENT

APPENDIX G

The NAEP Scale Anchoring Process for the 1992 Reading Assessment

Eugene G. Johnson, Ina V.S. Mullis, Jay R. Campbell, and Steven P. Isham

Educational Testing Service

Introduction

Beginning with the 1984 assessments, NAEP has generally reported students' subject area proficiency on 0-to-500 scales. These scales are used to report achievement for students at the various grades or ages assessed, including differences between performance from assessment to assessment for the nation and for various subpopulations of interest. To date, NAEP has used item response theory techniques to develop proficiency scales for reading, mathematics, science, writing, U.S. history, and civics.

Although average proficiency is an efficient summary measure, some of the most interesting NAEP results are those based on performance differences for different points in the scale distributions. To provide an interpretation for both the average results (What does a 306 on the 0-to-500 scale actually mean?) and changes in performance distributions (What does it mean that fewer students are reaching level 250?), NAEP invented a scale anchoring process to describe the characteristics of student performance at various levels along the scales—typically, at levels 200, 250, 300, and 350. The descriptions of student performance are presented in the reports accompanied by the percentages of students performing at or above the various scale levels.

Scale anchoring is a way of attaching meaning to a scale. Traditionally, meaning has been attached to educational scales by norm-referencing, that is, by comparing students at a particular scale level to other students. In contrast, the NAEP scale anchoring is accomplished by describing what students at selected levels know and can do.

On February 15-17, 1993, ETS applied a modified anchoring procedure to the 1992 reading achievement levels. As applied to the achievement levels, the anchoring process was designed to determine the sets of questions that students scoring at or above each achievement level cutpoint could perform with a high degree of success. A committee of reading experts, educators, and others was assembled to review the questions and, using their knowledge of reading and student performance, to generalize from the questions to descriptions of the types of skills exhibited at each achievement level.

The Scale Anchoring Analysis

A question was identified as anchoring at an achievement level for a given grade if it was answered correctly by at least 65 percent of the students in that grade scoring at the cutpoint of that achievement level, and by less than 65 percent of the students scoring at the cutpoints for any lower achievement level. In order to maximize the number of questions offered for consideration, the traditional discrimination criterion, which required that the chances of success at the next lower level be at least 30 percentage points lower, was not used.

To provide a sufficient pool of respondents in identifying anchor items, students at the cutpoint of each achievement level were defined as those whose estimated reading proficiency (as defined by their first composite plausible value) was within 12.5 points of the achievement level cutpoint on the NAEP scale. (The derivation of achievement level cutpoints on the NAEP scale is described in Appendix F.) This is consistent with previous anchoring procedures and provides an empirical estimate of the performance of students scoring at the cutpoint. To provide stable estimates, the calculations of the chances of success on an item had to be based on at least 75 students in the cutpoint interval; this was reduced from the previous requirement of 100 students to accommodate the small number of students reaching the advanced level.

The 1992 reading scale anchoring analysis was based on the scaled composite proficiency results for fourth, eighth, and twelfth graders participating in the 1992 national assessment. As illustrated below, for each item in the NAEP assessment, ETS determined the weighted percentage and raw frequency for students at each of the achievement levels correctly answering the item. This was done separately for each of the grade levels at which the item was administered. For example, the data for each item were analyzed as shown in the following sample.

Sample Scale Anchoring Results

Achievement level	Basic	Proficient	Advanced
Weighted p-value	0.22	0.49	0.73
Raw frequency	282	386	93

It should be noted that the percentages of students answering the item correctly at each of the achievement levels differ from the proportion of students scoring above each achievement level and from the overall p-value for the total sample at any one grade level.

Because the extended constructed response items were scored on an ordered scale with four scoring levels (minimal, partial, essential, and extensive), the above procedure, which relies on the notion of a correct or an incorrect response to an item, was generalized. To fit into the anchoring framework, each extended constructed-response item was treated as three distinct items corresponding to scores of partial or better, essential or better, and extensive. These distinct items were then analyzed in the same manner as items scored as correct/incorrect.

Thus, for example, an extended constructed-response item might anchor at the proficient level for partial or better responses, and at the advanced level for essential or better responses.

Because it was the lowest level being defined, the basic level did not have to be analyzed in terms of the next lower level, but only for the percentage of students at that level answering the item correctly. More specifically, for an item to anchor at the basic level:

1) The p-value for students at the basic level had to be greater than or equal to 0.65, and

2) the calculation of the p-value at that level had to have been based on at least 75 students to ensure adequate stability of the estimate of the p-value.

As an example, the following results are for an item anchoring at the basic level:

Basic Level Anchor Item Results

Achievement level	Basic	Proficient	Advanced
Weighted p-value	0.68	0.78	0.90
Raw frequency	308	413	115

For an item to anchor at the remaining levels, three criteria had to be met. For example, to anchor at the proficient level:

1) The p-value for students at the proficient level had to be greater than or equal to 0.65;

2) the p-value for students at the basic level had to be less than 0.65; and

3) the calculations of the p-values at both levels had to have been based on at least 75 students.

The following data set illustrates the results for a proficient level anchor item:

Proficient Level Anchor Item Results

Achievement level	Basic	Proficient	Advanced
Weighted p-value	0.34	0.73	0.95
Raw frequency	369	433	131

BEST COPY AVAILABLE

The same principles were used to identify anchor items at the advanced level. For example, the following results were obtained for an item anchoring at the advanced level:

Advanced Level Anchor Item Results

Achievement level	Basic	Proficient	Advanced
Weighted p-value	0.13	0.41	0.84
Raw frequency	313	423	106

By anchoring the achievement level cutpoints, instead of the entire interval, it is possible to determine the types of skills exhibited by all students within an interval. Thus, an item anchoring at the basic level cutpoint will be answered correctly by at least 65 percent of minimally basic students and will be answered correctly by at least that percentage of students in the basic interval. Since the NAEP results are reported in terms of the percentages of students at or above each of the cutpoints, it is important to be able to say what all students in the interval are likely to be able to do. In contrast, an anchoring procedure based on the interval identifies skills that a typical member of the interval (e.g., a typical basic student) likely possesses. While we could infer what a typical student in the basic interval can likely do, we would not be able to infer the skills of a minimally basic student.

A description of an entire achievement level interval can be inferred by comparing the descriptions for adjacent cutpoints. Thus, the description for the basic cutpoint tells what all basic students are likely to be able to do with increasing certainty as their reading proficiency increases. The description of the proficient cutpoint refers to the abilities of minimally proficient students, but also provides information about the capabilities of basic students scoring at the top of the basic interval. To extend the description of the advanced achievement level, since that interval does not have an upper boundary, an additional set of questions were identified as "almost anchoring" at the advanced level. These questions had probabilities of success between 50 and 65 percent for minimally advanced students and identify the types of skills that more advanced students are likely to possess.

For example, the results below are for an item almost anchoring at the advanced level:

Almost Advanced Level Item Results

Achievement level	Basic	Proficient	Advanced
Weighted p-value	0.11	0.31	0.55
Raw frequency	298	443	104

Preparing for the Reading Item Anchoring Panel Meeting

Table G-1 provides a breakdown of the numbers of anchored and almost anchored dichotomous items (i.e., items scored correct/incorrect) by content area and grade. The vast majority of these items anchored at some achievement level, or almost anchored at the advanced level. The remaining items that did not anchor were generally quite difficult.

Table G-2 provides similar information for the extended constructed response items that were scored on a four-point scale. As described above, each of these items was treated as three distinct items, corresponding to scores of partial or better, essential or better, and extensive. The counts in Table G-2 are in terms of these item parts. The item parts that did not anchor correspond to scores of extensive, and sometimes, essential or better.

In preparation for use by the scale anchoring panelists, the items were placed in notebooks by grade in the following order: anchored at basic, anchored at proficient, anchored at advanced, and almost anchored at advanced (chance of success between 50 and 65 percent at the advanced level). For cross-referencing purposes, the remaining items in the assessment were also included in the notebook under the "did not anchor" heading. (These were the items answered correctly by fewer than 50 percent of the students at the advanced cutpoint.) Each item was accompanied by its scoring guide (for constructed-response items), the chance of success on the item for students at each achievement level, the counts and weighted proportions of students at each level, the overall percent correct on the item for the total population of respondents, and the reading purpose and stance classifications for the item.

The anchoring process was further informed by results using the item mapping procedure. Item mapping provides additional information about the performance of students within each of the achievement level intervals, and of students who performed below the basic level. In item mapping, the items are arranged in the order of the proficiency level corresponding to a defined expected probability of success based on the item response theory parameters. The items, or short descriptions, are then displayed, along with the proficiency value associate with the selected probability of success. For consistency with the anchoring process, a .65 expected probability of success was used.

The Process for Developing the Descriptions

Twenty reading education experts participated in a three-day anchoring meeting. They represented teachers of the three grade levels, college professors, state curriculum supervisors, and researchers. (See Figure G-1 for a list of the participants.) The panelists were divided into three groups, one for each grade level. The grade-level groups worked independently for the most part, with periodic meetings across the three groups to reconcile views. With the framework for the 1992 reading assessment and the achievement level descriptions as a reference, panelists were asked to use the information in the anchor item notebooks and from the item mapping to describe the knowledge, skills, and reasoning abilities demonstrated by the students at the cutpoint of each achievement level. In addition, performance as depicted by the maps or items that almost anchored was taken as indicating beginning or emerging skills for students in the interval. Based on the items anchoring at each level and the item maps, the

Table G-1

Counts of Dichotomous Reading Items Anchoring by Content Area And Grade

Content Area	Basic	Proficient	Advanced	Almost Anchored at Advanced	Did Not Anchor
GRADE 4:					
Literary	15	12	9	2	1
Informational	13	10	10	1	4
GRADE 8:					
Literary	10	12	6	0	4
Informational	16	15	5	5	5
Task-oriented	8	17	3	2	3
GRADE 12:					
Literary	4	14	9	0	3
Informational	20	17	11	3	3
Task-oriented	17	8	6	2	3

Table G-2

Counts of Polytomous Item Parts* Anchoring by Content Area And Grade

Content Area	Basic	Proficient	Advanced	Almost Anchored at Advanced	Did Not Anchor
GRADE 4:					
Literary	1	1	2	2	5
Informational	2	1	2	2	5
GRADE 8:					
Literary	0	2	1	2	4
Informational	1	4	3	2	8
Task-oriented	2	0	2	1	3
GRADE 12:					
Literary	1	4	2	0	2
Informational	6	4	5	1	7
Task-oriented	1	1	3	0	3

---

* Each polytomous item was treated as three separate items corresponding to scores of partial or better, essential or better, and extensive.

panelists were asked to draft a description of achievement at each level. In drafting these descriptions, the panelists were instructed to consider the context of the assessment and to not overinfer skills from limited numbers of items. The draft descriptions were checked by staff against the anchoring data, edited, and sent to the panelists for final review. The final draft of the descriptions is presented in Figure G-2. Each achievement level at each grade corresponds to a cutpoint on the NAEP scale as described in Appendix F.

Figure G-1

Reading Scale Anchoring Panel

Eileen Baldwin
Trenton School District
Trenton, NJ

Margo Brill-Wigant
Department of Education
Santa Fe, NM

Miriam Chaplin
NCTE
Cherry Hill, NJ

Karen Costello
Connecticut State Dept. of Ed.
Hartford, CT

Eunice Greer
University of Illinois
Champagne, IL

Robert Harrison
English/Language Arts Coordinator
Charleston, WV

Diane Hoffman
Middle School Teacher
Sykesville, MD

Janet Jones
William D. Wade Elementary
Waldorf, MD

Barbara Kapinus
Council of Chief State School Officers
Washington, DC

Judith Langer
SUNY-Albany
Albany, NY

Patricia McConigal
5th Grade Teacher
Underhill, VT

Jim Martin-Rehrman
Westfield State College
Westfield, MA

Leslie Morrow-Mandel
Rutgers University
Scotch Plains, NJ

Susan Neuman
Temple University
Philadelphia, PA

Charles Peters
Oakland Schools
Waterford, MI

Gary Rice
Louisiana State University
Baton Rouge, LA

Timothy Shanahan
University of Illinois at Chicago
Chicago, IL

Robert Swartz
University of Massachusetts
Newtonville, MA

Gwendolyn Williams
K - 8 Reading Specialist
Landover, MD

Figure G-2

Anchor Descriptions of the Reading Achievement Levels

Grade 4 students ...

Basic (212)	... understand uncomplicated narratives and high-interest informative texts, identify an obvious theme, locate explicit information, summarize parts of text, and evaluate characters' actions.

Fourth-grade students at the basic level can read uncomplicated narratives with understanding. The *literary* texts at this level include fables and realistic fiction about familiar topics. These students can answer questions that focus on specific parts of the story. They are able to identify an obvious theme or message. They can take the perspective of characters that are familiar or similar to themselves and compare characters to each other. In addition, they can relate to the feelings of familiar characters, as well as interpret and evaluate the characters' actions.

Students at the basic level are able to gain information from high-interest *informative* texts. These students are successful when texts are structured as narratives and deal with relatively familiar topics. Students can search for and locate explicit information within the text, as well as provide evidence of straightforward comprehension of the text. They are able to select relevant information in order to provide a summarization focusing on part of the text. They can understand the sequence of events and identify situations described in the text. They can build simple inferences based on specific information. These students also are able to construct their own simple questions related to the passage.

Grade 4 students ...

Proficient (243)	... understand and interpret less familiar texts, provide textual support for interpretations, generalize across text, identify relevant information, understand subtleties in aspects of a story, relate text to background experiences, and formulate simple questions.

Fourth-grade students at the proficient level can form an understanding and extend the meaning of more difficult, unfamiliar *literary* pieces—those in culturally different or historical settings. They are able to respond to questions that require some interpretation. Some can construct responses to the story as a whole, as well as consider subtleties in aspects of the story. When given interpretations of the story, they can provide some justification and support for those interpretations. They are able to recognize multiple perspectives. In addition, they have the ability to connect information in the story to the author's purpose, as well as consider alternate possibilities for the story's development.

Students at the proficient level are able to gain information and to interpret the meaning of *informative* text that contains narrative elements and direct quotes. Their responses to increasingly more challenging questions provide evidence that they can search for, locate, select, prioritize, and apply relevant information. They can generalize across parts of the text. They

Figure G-2 (continued)

Anchor Descriptions of the Reading Achievement Levels

can relate information from the selection to their own background experiences and to inferences that are provided for them. They also are able to recognize an author's basic organizational pattern.

Grade 4 students ...

Advanced (275)	... interpret and examine meaning of text, summarize information across whole text, develop their own ideas about textual information, understand some literary devices, and begin to formulate more complex questions about text.

Fourth-grade students at the advanced level can form an understanding of what they read and extend, elaborate, and examine the meaning of *literary* texts. They can construct responses to a story by selecting relevant information and building their own interpretations that remain consistent with the text. They are able to summarize information across the whole story. They understand some literary devices, such as figurative language, and can interpret the author's intentions.

Students at the advanced level can gain information from what they read and can extend, elaborate, and examine the meaning of *informative* texts about less familiar topics. They are able to read for the purpose of gaining a more thorough understanding of a particular topic, and some can develop their own ideas based on the information presented in the passage. They can discriminate the relative importance of ideas in the text and are beginning to form more complex questions about the selection. They are able to provide an explanation of the author's techniques for presenting information.

Grade 8 students ...

Basic (244)	... understand familiar genres, recognize central theme or topic, identify the central purpose of practical documents, identify literal information, interpret and describe character traits, and connect information from across text.

Eighth-grade students' responses at the basic level demonstrate fundamental understandings of *literary* texts from familiar genres. These texts are not complex or abstract—they contain a single perspective and a central focus. These students can answer questions that focus on surface or literal understandings of the story. They can identify the basic theme of a story and can connect ideas within one section or across larger parts of the text. They are able to interpret and describe character traits.

Students' responses at the basic level demonstrate an ability to make concrete interpretations from *informative* texts (i.e., biographies, articles, informative narratives) that present information in a relatively straightforward manner. These students can recognize the central purpose by interpreting information across a text and by using structural text features, such as subheadings, exemplification, and organizational patterns. They are able to locate and to recognize explicitly stated information, as well as to connect information in one section of text with that from other sections. They are able to recognize the reasons an author might include partial information.

Students at the basic level are able to locate guidelines or directions that are explicitly stated in practical *documents*. They demonstrate some familiarity with documents, as well as an understanding of their purpose and usefulness. They can connect information presented within one section of a text to information in another section. They can articulate a personal view or choice about a document and support their opinion. In addition, they can use explicit directions to produce a specific textual form or document type.

Grade 8 students ...

Proficient (283)	... move beyond surface understanding of a text or multiple texts, make inferences about characters and themes, link generalizations to specific details, support an opinion about text, recognize an author's intentions, and use a document to solve simple problems.

Eighth-grade students at the proficient level are able to move beyond surface understandings of *literary* texts (i.e., historical fiction, tales) to develop fuller interpretations. They can recognize and interpret overall messages or themes implied in a literary piece. They are able to connect and make inferences about essential elements of stories and characters. They are able to interpret a character's ideas and feelings based on the events in the story and their own interpretation of the character's personality and role. These students can develop a perspective on a character's motivation by relying on their own understanding of human nature and essential story features, such as plot, dialogue, and description. They also can recognize an author's intentions and identify an author's use of symbolism to convey a story theme.

Proficient readers are able to locate and integrate information from different sections of an *informative* text and across multiple texts. At this level, students are able to gain information from textbook chapters, as well as biographies, articles, and informative narratives. These students can recognize a generalization and link it to specific details within the text. They demonstrate the ability to compare and contrast, as well as summarize information from across the text. They are able to form personal opinions about the content and provide supportive examples from text. They demonstrate an ability to use knowledge of organizational structures to gain information.

Readers at the proficient level are able to use multiple sources (i.e., time tables, instructions, maps) to locate information explicitly stated in a *document*. They can interpret the meaning of graphic symbols, such as map legends. They show the ability to perform tasks that

Figure G-2 (continued)

Anchor Descriptions of the Reading Achievement Levels

involve extracting information embedded within a document. They are able to discriminate among similar sources in accessing information to perform a task and solve a simple problem. They can understand how and why authors use text features and the relationship among particular features within documents, such as illustrations and examples.

Grade 8 students ...

Advanced (328)	**... compare and contrast information across multiple texts, connect inferences with themes, understand underlying meanings, integrate prior knowledge with text interpretations, and demonstrate some ability to evaluate the limitations of documents.**

Eighth-grade students reading at the advanced level are able to extend *literary* interpretations by relating personal knowledge to story characters and events. They demonstrate an understanding of fairly abstract themes and provide personal reactions to overall themes. They are able to interpret underlying meanings and complexities of characterizations and plot developments. They are able to connect inferences about characters' motives and feelings with story themes and provide supporting evidence from the story. In addition, they can relate themes across genres and to real-world situations. They also demonstrate the ability to consider the author's use of literary devices and relate it to an underlying theme.

Advanced eighth-grade readers are able to understand, to interpret, and to evaluate information presented in *informative* text. They are able to compare and contrast information within a text and across multiple texts and various genres. They make use of illustrations to enhance their interpretations of text. They can locate specific information embedded within text. They draw on knowledge from other subject areas and take a historical perspective in developing interpretations about text information. These students demonstrate the ability to formulate opinions about the information they read and support their ideas with appropriate text-based evidence.

Eighth-grade students at the advanced level are able to locate and to use very specific, highly embedded information in a fairly complex *document*. They use multiple pieces of information from various locations within a document to complete a task or solve a real-world problem. Many are able to evaluate the presentation of information in a document, recognize its limitations, and suggest improvements.

Grade 12 students...

Basic (269)	**... develop interpretations from a variety of texts, understand overall arguments, recognize explicit aspects of plot and characters, support global generalizations, respond personally to texts, use major document features to solve real-world problems.**

Figure G-2 (continued)

Anchor Descriptions of the Reading Achievement Levels

Twelfth-grade students at the basic level can gain meaning and develop interpretations from a variety of *literary* works (i.e., first-person adventures, narrative poems, tales). They respond to literature in a straightforward manner and focus their interpretations on specific aspects of a story. They are able to recognize fairly explicit aspects of plot development and characterization. Students at this level demonstrate surface understanding of characters' motives and are able to understand and use dialogue in constructing meaning. They can focus their attention, gain meaning, and develop interpretations from a character's perspective as well as their own. They respond personally to particular portions of a piece and report their responses to textual evidence.

Students at the basic level are able to gain information and to understand specific issues as a result of reading a variety of *informative* texts (i.e., encyclopedia entries, journal accounts, textbook chapters, science periodicals, editorials, and biographical essays). Students can gain information from reading individual texts or multiple texts on the same topic. They are able to recognize general arguments and viewpoints. They can use information from across text segments to make and support global generalizations. They are able to recognize explicitly stated problems and their solutions, as well as important causal relationships. In addition, they demonstrate an understanding of the potential contribution of illustrations and captions to readers' comprehension and engagement. These students are able to evaluate the importance of a particular issue and formulate an opinion.

Twelfth-grade students reading at the basic level are able to respond to forms, schedules, and practical *documents* adhering to most directions or guidelines. Drawing on text clues, they recognize and are able to locate explicit information stated in a document. These students demonstrate an understanding of the use of labels to group ideas and mark sections within documents. They are able to infer the purpose for document guidelines and compare a task completed according to the guidelines with another related task. In addition, these students are able to use accompanying maps, legends, symbols, and timetables to solve real-world problems. Students at the basic level recognize the most obvious limitations of a document's applicability and present personal reactions in response to document information.

Grade 12 students...

Proficient (304)	...integrate background experiences and knowledge with meaning from a variety of texts, interpret characters' motives, consider differing points of view, interpret literary devices, identify text structure and writing style, and apply document information to solve complex problems.

Proficient readers are able to form interpretations and express overall responses to *literary* texts (i.e., first-person adventures, narrative poems, tales). Drawing on their personal knowledge, they can interpret characters' motives and feelings, perceive significant character traits, identify similarities between characters, and develop an understanding of evolving characterizations within a story. In addition, they are able to find textual evidence to support their assumptions about characters and their actions. By delving beneath surface language and events, proficient readers are able to develop an understanding of the underlying intentions and communicative intent of dialogue. These readers integrate personal experiences with narrative or poetic elements and bring their real-world perceptions of the human condition to their literary interpretations. They are able to interpret figurative language and the symbolism suggested by major story elements.

Proficient readers are able to gain and to interpret relevant information from an individual *informative* passage or across multiple passages (i.e., encyclopedia entries, journal accounts, textbook chapters, science periodicals, editorials, and biographical essays). They are able to consider differing points of view in developing an understanding of text. They recognize the contributions of various texts in gaining overall understanding of a particular topic and are able to evaluate the credibility of different sources. Proficient readers demonstrate familiarity with informative genres by identifying organizational forms and recognizing patterns in writing style used by the author. They also are able to draw on background knowledge to interpret textual information and determine text reliability. Their responses to this type of text demonstrate an ability to analyze and make judgements about informative material.

Readers at the proficient level demonstrate comprehension of moderately complex and specific instructions presented in practical *documents*, including forms and schedules. Their responses demonstrate a clear understanding of a document's purpose. They are able to search documents to locate specific information from major sections and highly embedded details. They exhibit strategies for extracting and applying document information in successfully completing a multistep task. These readers are able to suggest alternative approaches to task completion and make choices based on an appropriate interpretation of the document's main features. They are able to access and use tabular and graphic information in making generalizations and decisions about real-world problems. They understand the purpose of a particular document and are able to tell the importance of complying with the guidelines.

Grade 12 students...

Advanced (348)	... construct complex understandings of multiple genres, interpret multidimensional aspects of characters, connect discipline-specific knowledge to text, examine author's craft, judge the value of informative sources, and suggest improvements for documents.

Advanced students are able to construct more complex and abstract understandings of *literary* texts by integrating personal knowledge and experiences with textual ideas and events. They are able to connect ideas and to relate interpretations across multiple types of literary genres. They are able to interpret the significance of major story elements, as well as draw on underlying meaning to develop a thorough understanding of an abstract theme. They consider non-explicit implications of language and dialogue within a literary piece. Drawing on their knowledge of human nature, they are able to interpret and describe nuances and multidimensional aspects of character relationships, feelings, and motives. They demonstrate an ability to examine their own personal understandings based on considerations of text meaning and real-world issues. They make use of their familiarity with literary elements to develop in-depth interpretations and examine critically the author's style and use of literary devices.

Students reading at the advanced level demonstrate the ability to synthesize and critically examine information presented in individual and multiple *informative* texts. They use information presented within a text to build overall understandings of conditions occurring across time. These readers cad identify the significance of events and draw on general background experiences, as well as discipline-specific knowledge to advance their understanding of information presented within text. They use genre-appropriate strategies to glean specific information, search for evidence to support generalizations, evaluate the credibility of multiple sources and identify potentially different uses for information gained from different sources. They perceive ways in which a point of view is expressed in an author's language and make judgements about the author's intent. By considering a text's purpose, structure, and content they are able to make and support judgements about its informative value.

Advanced readers demonstrate an ability to manage various organizational structures in accessing and applying information presented in documents, including forms and schedules. They are able to use specified directions and guidelines to complete highly detailed tasks. In addition, they are able to integrate text with graphic organizers in interpreting the meaning of written directions. These students are able to follow a series of complex steps specified by document directions in order to extract relevant information for a particular purpose. Based on a thorough examination of document text and structure, they make thoughtful and appropriate recommendations for improving the usefulness and presentation of information within a document.

REFERENCES CITED IN TEXT

REFERENCES CITED IN TEXT

Abt Associates. (1991). *Prospects: The National Longitudinal Study of Chapter I children* (Final progress report for design contract No. LC89027001). Chicago, IL: Author.

Andersen, E. B. (1980). Comparing latent distributions. *Psychometrika, 45*, 121-134.

Anderson, R. C., Hiebert, E. H., Scott, J. A., and Wilkinson, I. A. G. (1984). *Becoming a nation of readers: The report of the Commission on Reading*. (U.S. Department of Education: The National Institute of Education).

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508 - 600). Washington, DC: American Council on Education.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17*, 191-204.

Beaton, A. E., & Johnson, E. G. (1992). Overview of the scaling methodology used in the National Assessment. *Journal of Educational Measurement, 26(2)*, 163-175.

Beaton, A. E., & Johnson, E. G. (1990). The average response method of scaling. *Journal of Educational Statistics, 15*, 9-38.

Beaton, A. E., & Zwick, R. (1990). *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly.* (No. 17-TR-21) Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Bourque, M. L., & Garrison, H. H. (1991). *The levels of mathematics achievement. Vol. I, national and state summaries.* Washington, DC: National Assessment Governing Board.

Burke, J., Braden, J., Hansen, M., Lago, J., Tepping, B. (1987). *National Assessment of Educational Progress -- 17th year sampling and weighting procedures. Final report.* Rockville, MD: Westat, Inc.

Cochran, W. G. (1977). *Sampling techniques.* New York, NY: John Wiley & Sons.

Curry, L. (1987, April). *Group decision process in setting cut-off scores.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Dole, J. A., Duffy, G. G., Roehler, L. R., and Pearson, P. D., (1991). Moving from the old to the new: Research in reading comprehension instruction. *Review of Educational Research, 61.*

Educational Testing Service (1987). *ETS standards for quality and fairness.* Princeton, NJ: Author.

Engelen, R. J. H. (1987). *Semiparametric estimation in the Rasch model.* Research Report 87-1. Twente, the Netherlands: Department of Education, University of Twente.

Fitzpatrick, A. R. (1989). Social influences in standard-setting: The effects of social interaction on group judgments. *Review of Educational Research, 59,* 315-328.

Friedman, C. B., & Ho, K. T. (1990, April). *Interjudge consensus and intrajudge consistency: Is it possible to have both on standard setting?* Paper presented at the annual meeting of the National Council for Measurement in Education, Boston, MA.

Guthrie, J. T., & Greaney, V. (1991). Literacy acts. In R. Barr, M. Kamil, P. Mosenthal, & P.D. Pearson (Eds.), *Handbook of reading research: Volume II.* (New York, NY: Longman).

Hambleton, R. K., & Bourque, M. L. (1991). *The levels of mathematics achievement. Vol. II, technical report.* Washington, DC: National Assessment Governing Board.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis.* New York, NY: John Wiley & Sons.

Hoijtink, H. (1991). *Estimating the parameters of linear models with a latent dependent variable by nonparametric maximum likelihood.* Research Bulletin HB-91-1040-EX. Groningen, The Netherlands: Psychological Institute, University of Groningen.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity.* Hillsdale, NJ: Erlbaum.

Jerry, L. (1993). *The NAEP computer-generated reporting system for the 1992 Trial State Assessments.* Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Johnson, E. G., & Allen, N. L. (1992). *The NAEP 1990 technical report* (No. 21-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Johnson, E. G., Mazzeo, J., & Kline, D. L. (1993). *Technical report of the NAEP 1992 trial state assessment program in mathematics.* (No. 23-ST05) Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Johnson, E. G., & Rust, K. F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17*, 175-190.

Johnson, E. G., & Zwick, R. (1990). *Focusing the new design: The NAEP 1988 technical report* (No. 19-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Keyfitz, N. (1951). Sampling with probability proportional to size; adjustment for changes in probabilities. *Journal of the American Statistical Association, 46*, 105-109.

Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association, 73*, 805-811.

Langer, J. A. (1989). *The process of understanding literature*. (Technical report No. 2.1.) Albany: State University of New York, Center for the Learning and Teaching of Literature.

Langer, J. A. (1990). The process of understanding: Reading for literary and informative purposes. *Research in the Teaching of English, 24*, 229-257.

Lindsey, B., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association, 86*, 96-107.

Little, R. J. A., & Rubin, D. B. (1983). On jointly estimating parameters and missing data. *American Statistician, 37*, 218-220.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: John Wiley & Sons.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mazzeo, J. (1991). Data analysis and scaling. In S. L. Koffler, *The technical report of NAEP's 1990 Trial State Assessment program* (No. ST-21-01). Washington, DC: National Center for Education Statistics.

Mazzeo, J., Chang, H., Kulick, E., Fong, Y. F., & Grima, A. Data analysis and scaling for the 1992 Trial State Assessment in mathematics. In E. G. Johnson, J. Mazzeo, & D. L. Kline, *Technical report of the NAEP 1992 Trial State Assessment program in mathematics*. Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Mazzeo, J., Johnson, E. G., Bowker, D., & Fong, Y. F. (1992). *The use of collateral information in proficiency estimation for the Trial State Assessment*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196.

Mislevy, R. J. (1990). Scaling procedures. In E.G. Johnson and R. Zwick, *Focusing the new design: The NAEP 1988 technical report* (No. 19-TR-20). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*, 993-997.

Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17(2)*, 131-154.

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. In A.E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report* (No. 15-TR-20). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Mislevy, R. J., & Stocking, M. L. (1987). *A consumer's guide to LOGIST and BILOG.* (ETS Research Report 87-43). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Wu, P-K. (1988). *Inferring examinee ability when some item responses are missing* (ETS Research Report RR-88-48-ONR). Princeton, NJ: Educational Testing Service.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16(2)*, 159-176.

Muraki, E., & Bock, R. D. (1991). *PARSCALE: Parameter scaling of rating data.* Chicago, IL: Scientific Software, Inc.

National Assessment Governing Board (1989). *Setting achievement goals on NAEP, a draft policy statement.* Washington, DC: Author.

National Assessment of Educational Progress (1992). *1992 policy information framework.* Princeton, NJ: Educational Testing Service.

Petersen, N. (1988). *DIF procedures for use in statistical analysis.* Internal memorandum.

Potter, F. (1988). Survey of procedures used to control extreme sampling weights. *Proceedings of the Section on Survey Research Methods* (pp. 453-458). Washington, DC: American Statistical Association.

Rogers, A. M. (1991). *NAEP-MGROUP: Enhanced version of Sheehan's software for the estimation of group effects in multivariate models* [Computer program]. Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1991). EM and beyond. *Psychometrika, 56*, 241-254.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Rust, K. R., & Bryant, E. (1991). *The IEA Reading and Literacy Study design and implementation: National and international perspectives, population definitions and sample design.* Presented at the annual meeting of the American Educational Research Association, Chicago, Illinois.

Rust, K. R., Burke, J., Fahimi, M., & Wallace, L. (1992). *1990 National Assessment of Educational Progress sampling and weighting procedures. Part 2 - National Assessment.* Rockville, MD: Westat, Inc.

Sheehan, K. M. (1985). *M-GROUP: Estimation of group effects in multivariate models* [Computer program] Princeton, NJ: Educational Testing Service.

Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. *Journal of Educational Measurement, 25*, 259-274.

Stokes, L. (1990). A comparison of truncation and shrinking of sample weights. *Proceedings of the Annual Research Conference* (pp. 463-471). Washington, DC: U.S. Bureau of the Census.

Stone, C. A., Andenmann, R. D., Lane, S., & Liu, M. (1993). *Scaling QUASAR's performance assessments.* Paper presented at the annual meeting of the American Educational Research Association.

Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association, 82*, 528-550.

Thomas, N. (1992). *Higher order asymptotic corrections applied in an EM algorithm for estimating educational proficiencies.* Unpublished manuscript.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley Publishing Co.

Wainer, H. (1974). The suspended rootogram and other visual displays: An empirical validation. *The American Statistician, 28(4)*, 143-145.

Wingersky, M., Kaplan, B. A., & Beaton, A. E. (1987). Joint estimation procedures. In A. E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report*. (No 15-TR-20) Princeton, NJ: National Association of Educational Progress, Educational Testing Service.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17(2)*, 155-173.

Zieky, M. (1993). Practical questions in the use of DIF statistics. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Zwinderman, A. H. (1991). Logistic regression Rasch models. *Psychometrika, 56*, 589-600.

☆ U.S. GOVERNMENT PRINTING OFFICE:1994-300-807/13106

ISBN 0-16-043109-3

90000

9 780160 431098

BEST COPY AVAILABLE

DEPARTMENT OF EDUCATION · UNITED STATES OF AMERICA

NCES 94-472

330

BEST COPY AVAILABLE